The Deep Extent Of Mental Autonomy

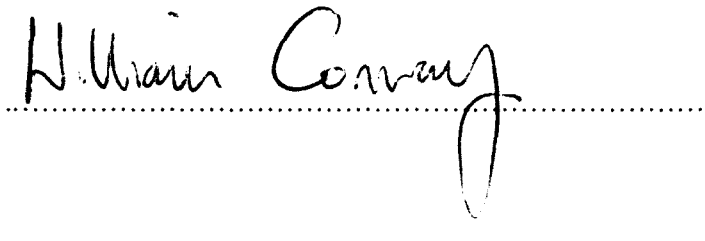
William Cassidy Stronach Conway

Degree of Doctor of Philosophy

The University of Edinburgh

1998

I hereby declare that I have composed this thesis, and that the work contained herein is entirely my own:



Contents

Acknowledgements

First of all, I would like to thank my main supervisor, Peter Lewis, for his helpful comments on an earlier draft of this thesis. His quiet weighing of linguistic facts has prevented me from over-stating certain points and from under-stating others. I would also like to thank my second supervisor, Dr. Alexander Bird, for his excellent and very thorough comments on chapter 2. In writing this thesis, I have benefited from discussions with various people. In particular, I have benefited from discussions with Maura Tumulty, now at the University of Pittsburgh, without whom I would never have seen the distinctively human side to Wittgenstein's *Philosophical Investigations*. Special thanks are also due to my good friend and philosophical sparring partner, Michael Higgins, with whom, over the past five or six years, I have had the pleasure to discuss some of the issues now contained in this thesis. His sharpness of intellect has been inspiring. I would also like to thank Pal Opdal, not only for his lessons in Norwegian, but more importantly for his friendship and much needed encouragement through the final stages of this work. But my greatest debt is to Suzanne, whose constant presence has prevented my spirits from flagging on more than one occasion. Finally, I am grateful for having had the opportunity to present earlier versions of chapters 4 and 6 at work-in-progress meetings in the University of Edinburgh, and I would like to express my gratitude to the Student Awards Agency for Scotland for the award of a three year Major Studentship grant which made this study possible.

Abstract

The central aim of this thesis is to argue that the autonomous nature of mentalistic explanation presents a stronger constraint on what counts as a satisfactory statement of the relation between the mental and the physical than can be acknowledged within the metaphysical framework of non-reductive physicalism. Although the chief merit of non-reductive physicalism appears to be its ability to respect the irreducibility of mental concepts to physical concepts, whilst respecting the primacy of the physical ontology, I claim that its commitment to the principles of physicalism prevents that framework from being able to accommodate what I will refer to as the deeper extent of the autonomous nature of mentalistic explanation. The deeper extent of the autonomous nature of mentalistic explanation manifests itself in the fact that the work carried out by mentalistic explanations is completely separate from the work carried out by physicalistic explanations. I claim that the deeper extent of the autonomous nature of mentalistic explanation cannot be recognised within a metaphysical framework which claims to recognise the primacy of the physical ontology because recognising deep autonomy requires giving up the assumption that the mental must be related to the physical in the manner appropriate to discharging such metaphysical principles.

I defend the claim that we can recognise the deeper extent of the autonomous nature of mentalistic explanation if we take our successful explanatory practices as the starting point of our investigation, and only then revert to the question of how best to articulate the relation between the mental and the physical. My claim is that there is an intrinsic connection between the nature of the mental and the nature of human relationships, and I therefore suggest that the autonomous nature of mentalistic explanation ought to be understood in connection with the autonomous nature of human relationships. The basic ideas in this thesis are derived by combining features of Wittgenstein's rule following considerations with features of John MacMurray's approach to human relationships. On the basis of this combination, I argue for the more specific claim that there is an intrinsic connection between what it means to say that an individual has the capacity to think and what it means to say that he has the capacity to be involved in various types of human relationships. This connection is then used to develop a non-causal account of human action to challenge the physicalist's causal account, which will be used to support the claim that mentalistic explanations are autonomous with respect to physicalistic explanations in the deeper sense.

I conclude by arguing that the considerations which put us in position to recognise the deeper extent of the autonomous nature of mentalistic explanation ought to constrain our statement of the relation between the mental and the physical, and I suggest that this statement should be consistent with the way in which mentalistic and physicalistic explanations carry out their work in our explanatory practices. I claim that individuals are subject to mentalistic explanations in so far as they have a life to live in the world with other people, and that individuals are subject to physicalistic explanations in so far as human beings are creatures whose life has a natural biological dimension. But rather than identify the mental with the physical, and thereby compromise the deeper extent of the autonomous nature of mentalistic explanation, I suggest that this relation might be understood in terms of the fact that the mental is embedded in the dimension of human life which is constituted by the involvement of individuals in various types of relationships with each other, and that the dimension of human life in which physicalistic explanations are operative is presupposed as the causal background which must be in place if individuals are to have such a life to live in the world.

Chapter 1: Introduction

Recent attempts to articulate the nature of the relation between the mental and the physical have been largely constrained by a sensitivity to the autonomous nature of mentalistic explanation. To those who are concerned to conduct their thinking within the parameters of the received scientific world-view, this means that a satisfactory statement of the relation between the mental and the physical not only has to be consistent with a recognition of the irreducibility of mental concepts to physical concepts, but also with a recognition of the primacy of the physical ontology. Consequently, the relation between the mental and the physical has been most commonly articulated within the metaphysical framework of non-reductive physicalism. Not only does non-reductive physicalism claim to display a bias toward the primacy of the physical ontology; it also claims to be sensitive to the autonomous nature of mentalistic explanation. However, there appears to be a tension at the heart of this position. On the one hand, the primacy of the physical ontology seems to imply that it must be possible to explain everything that exists and occurs in terms of what exists and occurs in the physical domain; but on the other hand, the irreducibility of mental concepts to physical concepts seems to imply that it is not possible to give an exhaustive characterisation of every aspect of human behaviour in terms drawn exclusively from the physical sciences.

It is not my aim to contribute to the arguments which purport to remove this tension, by explaining how the primacy of the physical ontology can be reconciled with the autonomous nature of mentalistic explanation. Rather is it my aim to argue that there is a deeper extent to the autonomous nature of mentalistic explanation than can be recognised when the mental is related to the physical in the manner suggested by the metaphysical framework of non-reductive physicalism. Whilst that framework claims to respect the autonomous nature of mentalistic explanation in respecting the irreducibility of mental concepts to physical concepts, it seems to me that this is not sufficient to capture the deeper extent to which mentalistic explanations are autonomous with respect to physicalistic explanations. I argue that mentalistic explanations are autonomous in the much deeper sense that they can carry out the work required of them without implicating the explanatory support of underlying physicalistic explanations. Consequently, I claim that a satisfactory statement of the relation between the mental and the physical is one which is sensitive to the deeper extent of mental

autonomy, and hence that this relation cannot be properly articulated within the metaphysical framework of non-reductive physicalism.

I think the main reason why non-reductive physicalism cannot recognise the deeper extent of the autonomy of mentalistic explanations lies with one of the central motivating factors for that position. Non-reductive physicalism is largely motivated by the need to account for the causal efficacy of mental events with respect to the behaviour they purport to explain. This is achieved through the imposition of a metaphysical ordering onto reality which secures the identification of mental events with physical events, and hence at the same time secures the causal efficacy of mental events with respect to the individual's behaviour. Here is where I think the problem lies: if mental events can be regarded as causally efficacious with respect to the individual's behaviour only in so far as they are identical with physical events, the result is that the mentalistic explanations which cite these mental events can be said to carry out the explanatory work required of them only in so far as they derive support from the explanatory resources of the physical sciences. But if this is the case, mentalistic explanations cannot be granted full autonomy with respect to physicalistic explanations. That can only be granted if we can find a way of explaining human action which does not require the identification of mental events with physical events, so that mentalistic explanations need not be said to implicate physicalistic explanations in order to carry out the explanatory work required of them.

A useful starting point for achieving this aim is to take our various explanatory practices as basic, and only once the role of mentalistic explanations and physicalistic explanations have been worked out within their appropriate practices do we revert to the question of the relation between the mental and the physical. The merit of taking this approach is that it straight away puts us into position to appreciate the actual role of mentalistic explanations, and it puts us into position to understand how mentalistic explanations can be said to fulfil that role without deriving explanatory support from physicalistic explanations. The recommendation is that instead of starting out with a full-blown metaphysical ordering of reality, we start out with an investigation into the nature of our explanatory practices as they stand. Our explanatory practices can then be our guide to our statement of the relation between the mental and the physical, rather than our statement of the relation between the mental and the physical being our guide to the nature of our explanatory practices. This starting point is recommended by Baker, who suggests that:

instead of beginning with a full-blown metaphysical picture, we should begin with a range of good explanations, scientific and commonsensical...Start with explanatory practices and let the metaphysics go. (1993: 95).

What I want to do is develop this starting point by arguing that the proper contexts in which to assess the work carried out by mentalistic explanations are the contexts created by our involvement in various types of human relationships. I want to argue on the strength of this that mentalistic explanations are autonomous with respect to physicalistic explanations in the deep sense that the work carried out by the former is completely separate from the work carried out by the latter, and I contend that this presents a strong constraint on what counts as a satisfactory statement of the relation between the mental and the physical. My central ideas are derived by combining aspects of Wittgenstein's treatment of rule governed practices with aspects of John MacMurray's treatment of human relationships. I will argue on the basis of this combination that there is an intrinsic connection between the nature of the mental and the nature of human relationships, and that the autonomous nature of mentalistic explanation is therefore tied to the autonomous nature of human relationships. Specifically, I will argue that there is an intrinsic connection between what it means to have the capacity to think and what it means to have the capacity to be involved in various types of relationships with other people; this will enable me to tie the success of mentalistic explanations in carrying out their work to their ability to satisfy the understanding we seek in our everyday relationships with each other. I will then proceed to develop the central features of this intrinsic connection into an account of human action which does not require the identification of mental and physical events, and I will eventually go on to suggest a means of articulating the relation between the mental and the physical based on these considerations.

Here is a brief outline of the following chapters. In the second chapter I will review various non-reductivist approaches to the autonomy of the mental, and I will explain why I think the deeper extent of the autonomy of the mental cannot be recognised within this framework. In the third chapter I will argue that the proper contexts in which to investigate the autonomous nature of mentalistic explanations are the contexts created by our involvement in various types of human relationships. The fourth chapter will be concerned with a detailed application of these ideas to what it means to say that an individual has the capacity to think, and I will draw out some connections between the physicalistic approach to the mental and its underlying conception of human relationships. In the fifth chapter I will be concerned to develop these ideas into a non-causal account of rational action. This will be used to

challenge the non-reductivist's assumption that the notion of causality must be built into our conception of what it means to act for a reason, which in turn will challenge the key assumption that rationalistic explanations must be a species of causal explanation, where the relevant causal processes take place at the physical level.

In the sixth chapter I will develop an account of the externalistic individuation of thoughts out of these arguments, which will support the non-casual approach to rationalistic explanations, and which will put pressure on the identity claim. In the seventh chapter I will discuss the relevance of the brain to my account of thinking, and I will argue that physicalistic approaches to thinking tend to over-inflate the importance of the brain to the extent that brain functioning is regarded as itself constituting thinking. In the eighth chapter, in order to round off my main argument, I will discuss the case of non-human animals to assess their claim to be treated as having a mind, which I seem to have ruled out by the claim that the nature of the mental is intrinsic to the nature of human relationships. Finally, in the ninth chapter, I will draw some conclusions concerning the constraints we face when attempting to articulate the nature of the relation between the mental and the physical.

Chapter 2: Physicalism and Autonomy

1.Introduction

It is possible to adopt two different modes of explaining human behaviour. Which mode of explanation is adopted depends on whether our interest lies in understanding others as rational agents, or as purely physical beings. Understanding others as rational agents requires us to explain their behaviour by drawing on the array of concepts which reveal them to be acting in light of reasons; understanding others as purely physical beings requires us to explain their behaviour by drawing on the array of concepts which reveals them to be part of the explanatory order of physical nature. The use of the latter concepts is restricted to investigations carried out within the field of the physical sciences, whereas the use of the former concepts is rather spontaneous and immediate within the contexts of our everyday lives. Since both modes of explanation must be in place if we are to have a satisfactory account of what it means to be a human being, it is natural to ask how we are to conceive the relationship that obtains between these modes of explanation, and in turn it is natural to ask whether our conception of this relationship imposes any constraints on what we ought to accept as a satisfactory statement of the relationship that obtains between the mental and the physical.

It is widely held that non-reductive physicalism offers the most plausible account of the nature of the relationship that obtains between the mental and the physical. The reason for this is simple. Non-reductive physicalism claims to recognise the fact that mental concepts are irreducible to physical concepts, whilst maintaining that the physical ontology is exhaustive of what there is. Which is to say that non-reductive physicalism claims to recognise the autonomous nature of mentalistic explanation whilst affirming the ontological primacy of the physical. This is its main attraction. Despite assurances to the contrary, however, it seems obvious to some that the irreducibility of mentalistic explanations to physicalistic explanations implies that the bias toward a monistic physical ontology ought to be reconsidered.[1] But a number of arguments have been developed to establish that no such

[1] Madell (1988a: 144), for instance, writes: "it seems more and more extraordinary that the view that reality consists wholly of agglomerations of elementary particles should nevertheless be seen as allowing us to talk of thought and feeling, reason, agency and value, truth and falsity. Only the huge

implication holds. The arguments to block this implication, which I shall come to presently, suggest that it is indeed possible to give a scientifically respectable account of the nature of the relationship between the mental and the physical, which nonetheless respects the autonomous nature of mentalistic explanation. To be successful, the non-reductive physicalist is therefore required to find a balance between the autonomous nature of mentalistic explanation and the ontological primacy of the physical.

Here is how this balance is achieved. Sensitivity to the autonomous nature of mentalistic explanation requires the physicalist to substantiate his ontological claims with an account of the nature of the dependence of the mental on the physical that does not immediately involve him in any kind of reductionist programme. Otherwise, there will seem little reason to continue thinking of physicalism as being able to provide a complete and exhaustive account of everything that exists, whilst at the same time being able to respect the autonomous nature of mentalistic explanation. The question of whether a bias toward the primacy of the physical ontology implies the reducibility of the mental to the physical is therefore the question of whether it is possible for the physicalist to find a dependence relation which does not also qualify as a reductive relation. So what is required for a comprehensive statement of the non-reductivist's position is not a purely ontological thesis, but this together with an account of the dependence of the mental on the physical. Hellman and Thompson put it thus:

> Although a purely ontological thesis is a necessary component of physicalism, it is insufficient in that it makes no appeal to the power of physical law...we seek to develop principles of physical determination that spell out rather precisely the underlying physicalist intuition that the physical facts determine all the facts. The goal then is to show that these principles do not imply reductionism. (1975: 552).

The question might be still more complicated than this, however, since the need to find principles of determination or dependence that do not imply reductionism only seems to be part of the physicalist's task. Arguably, the physicalist's task is complicated by the fact that even if he combines a purely ontological thesis with principles of dependence and determination, this does not yet seem sufficient to capture the sense in which the physical facts explain all the facts, or the sense in which everything which exists does so in virtue of the physical. Granted, this might seem to impose an extremely strong demand on the physicalist; but it is possible that a weaker demand, one which is satisfied with an ontological thesis together with principles of dependence and determination, still fails to discharge the

---

pressure toward some sort of monistic view could possibly explain the acceptance of a position so fundamentally implausible."

central commitments of the physicalist's explanatory programme. For the mere fact that mental properties are held to be dependent on physical properties (even if that dependence relation is so strong that it becomes possible to effect the nomological correlation of mental properties with physical properties) does nothing to explain how the physical ontology is primary, or to explain how mental properties are thereby embedded in the physical structure of the world.

Poland (1994: 16-17) insists that in addition to providing an ontological statement and principles of dependence, the physicalist must set out very definite principles which explain how mental properties are instantiated *in virtue of* the instantiation of physical properties, and this must involve a precise specification of those physical properties which are nomologically sufficient for, and relevant to, the instantiation of certain mental properties. As Poland sees it, the physicalist is obliged to explain how the instantiation of certain physical properties *realises* the instantiation of certain mental properties, and this means that he has to explain how the nature or essence of mental properties can be constituted by the instantiation of a specific configuration of physical properties. But this seems rather strong, since it is difficult to understand how the intrinsic nature of mental properties could be realised by a configuration of physical properties. How, for instance, can an individual's thinking about the hardships endured throughout last winter be constituted by a configuration of physical properties, when such configurations simply do not display the intentionality intrinsic to the individual's thinking? It does not help matters to refer to the fact that there are many different physical configurations capable of constituting the instantiation of the property transparency, as Poland claims, since transparency is a physical property, and as such it already fails to display the intentionality that cannot be constituted by instantiations of configurations of physical properties anyway. As Madell puts it, the problem is that:

> no series of physical events is intrinsically about anything; it is only as interpreted by human beings that such series can be seen as possible representations of processes of thought. That is to say, the intrinsic *directedness* of thought isn't something which one can meaningfully ascribe to a set of physical items or events. (1988b: 113-114).[2]

---

[2] But are there not cases in which it is correct to say that the property of intentionality is constituted by configurations of physical properties: plants growing toward sunlight and thermometers recording room temperatures seem to display intentionality, yet all we have here are particular configurations of biological and physico-chemical properties? Not so- it seems to me that the intentionality in such cases is, as Madell puts it, only as interpreted. The point might be made in terms of the fact that whereas we can read intentionality into the activity of plants and thermometeres from an external stand-point, their activity is not itself intentional because it is not carried out for *reasons had* by the plants and the thermometers.

Yet Poland insists that the physicalist has to be able to explain how mental properties are embedded in the physical structure of the world, otherwise the primacy of the physical ontology cannot be properly acknowledged. That is, a purely ontological statement, together with principles of determination and dependence, are simply not sufficient to discharge the physicalist's commitment to the primacy of the physical ontology. What forces these very strong demands on the physicalist is the need for physicalism to discharge its function as a programme for explanatory unification; it is the need for physicalism to provide a precise explanation of how mental properties are embedded in the physical structure of the world. However, it seems to me that we should acknowledge the limits to the explanatory resources of the physical sciences, rather than attempt to apply and reapply them beyond their own proper field of interest. For unless these limits are acknowledged, unless we respect the resistance of the mental to be incorporated into the physicalistic explanatory programme, it seems to me that the upshot will be time and effort misspent trying to force the mental into a system in which it is not going to fit.

In the remainder of this chapter my central concern will be whether the commitment to the primacy of the physicalistic ontology is in fact consistent with the recognition of the autonomous nature of mentalistic explanation. It will be suggested that this commitment forces the physicalist to articulate the relationship that obtains between the mental and the physical in such a way that threatens to undermine mental autonomy. What I hope to achieve out of this discussion is a motivated rejection of the metaphysical framework of non-reductive physicalism. My rejection of that framework will be not be based on the strong claim that it has been shown to be internally inconsistent, although I will certainly be testing that consistency at various points; rather will it be based on the weaker claim, that non-reductive physicalism cannot fully discharge its physicalist commitments without compromising the deeper extent of the autonomous nature of mentalistic explanation. I will begin with a discussion of some of the classic arguments for non-reductive physicalism, and then I will move on to consider some more recent developments in this field.

2.Arguments for non-reductive physicalism

2.1a.The Variable Realisability Of The Mental

It might seem natural to suppose that if individuals are purely physical beings, then there must be an explanation of every aspect of human behaviour in terms drawn from the physical sciences. If this is the case, then it does not seem possible to acknowledge mentality as a real and autonomous feature of our world. The worry seems to be that if we start out from the assumption that individuals are just complex physical systems, then we have to give up the hope of securing a degree of autonomy for the mental. But this does not seem to mesh very well with the fact that explanations of an individual's behaviour in mental terms are such that they cannot be replaced by, or even derived from, explanations of his behaviour in purely physical terms. What seems to be required, if we are to retain the basic commitment to physicalism, is a way of blocking the implication from the basicness of the physicalistic ontology to the exhaustiveness of its explanatory programme. It has been thought that this can be achieved if we can leave room for irreducible modes of explanation within the basic physicalistic framework.

Putnam (1975a) suggests that we might understand what it means to say that a mode of explanation is autonomous if we think of the following example: an explanation of the failure of the square peg to fit into the round hole is that the board and the peg are rigid and that the round hole is smaller than the square peg. This seems obvious, and let us face it, rather boring. But the point of the example is that the explanation at the level of everyday geometrical relations, as opposed to the level of particle physics, brings out the relevant structural features of the situation which enable us to understand why the peg does not fit into the hole on this and other occasions, when the same higher level structural features are present. In fact, the same explanation would hold again whether the peg was made of wood, rubber or steel, and whether this particular atom was positioned here, and whether that particular atom was positioned there. In this respect, the higher level functional explanation is autonomous with respect to the explanation at the lower level of particle physics.

Putnam's suggestion is that the key to understanding how the mode of explanation appropriate to the mental can retain its autonomy with respect to the mode of explanation appropriate to the physical is the notion of functional isomorphism. The basic idea is that

mental states are functional states which are individuated by their causal functional role. To conceive mental states as such is to rule out the possibility of explaining them in terms of lower level physical explanations, even if the latter apply unquestionably to the physical states which, as a matter of fact, happen to realise them on any given occasion. Two states can be functionally isomorphic, and in this sense be the same states, yet be realised by completely different physical states. So if mental states are individuated according to their functional role, their identity as the states they are must remain distinct from the identity of their physical realisation bases. They can retain their identity as the mental states they are, regardless of whether they are realised by cheese, cogs or copper wire. Of course, this is not to suggest that we really ought to entertain the possibility that human beings might be composed out of these things. It is only to drive home the point that it is impossible to derive an explanation of an individual's behaviour in mental terms from an explanation of his behaviour in physical terms. As far as mental states are concerned, it is quite irrelevant that they have the physical realisation they do. So the possibility of mental states being realised in a variety of different physical states makes the connection between mental and physical explanations accidental, such that explanations in mental terms cannot be deductively derived from explanations in physical terms. The result of this seems to be that:

> we do have the kind of autonomy that we are looking for in the mental realm. Whatever our mental functioning may be, there seems to be no serious reason to believe that it is explainable by our physics and chemistry. (1975a: 297).

Putnam's deeper point in all of this is that physicalism can be made out to be perfectly consistent with the recognition of the mental as a real and autonomous feature of our world, because the reduction of the mental to the physical depends on the possibility of identifying mental state types with physical state types. The underlying point here is that whereas reductionism is certainly an implication of type-identity physicalism, it is not an implication of token-identity physicalism. Token-identity physicalism can hold that every given mental state is realised in some physical state, whilst it is not necessary that every mental state type is identical with a physical state type. This is just as well, since it is a highly unrealistic and implausible requirement that whenever individuals, creatures or organisms have the very same mental state type in common, for example, being in pain, they must also have the very same physical state type in common. It is only if this claim is made good that it can be said that physicalism implies reductionism, and that the explanatory scope of the physical programme is wide enough to exhaustively incorporate the mental. But as Putnam points out, this would presuppose that it was possible to:

specify a physical chemical state such that any organism (not just a mammal) is in pain if an only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical chemical state. This means that the physical chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain...etc. At the same time, it must not be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extra-terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain. (1975b: 436).

The variable realisability of the mental comes to this: it is possible for any given mental state to be realised in physically diverse ways, as illustrated with reference to the variety of different species which may be said to be in pain, without necessarily having to be in the same physical-chemical states. Therefore, it is highly unlikely that mental state types can be nomologically linked to physical state types in the manner required if physicalism is to imply that mental explanations can be derived from physical explanations. But this cannot give us an adequate conception of the mental since mental states and events are individuated according to rationalistic and justificatory relations which cannot be reproduced by fixing the location of physical states and events within any type of causal network. Regardless of how complex or how tightly interlocking these physical states and events happen to be, they will remain in brute contingent connections, and as such their network will fail to capture the notions of 'must' and 'ought' which figure constitutively in the network of rationalistic and justificatory relations definitive of intentional mental states and events.

Thus, it is rather questionable to offer an account of the autonomy of the mental in terms of the failure of explanations at the level of physics to imply functional explanations, since this does not seem to touch on the question of the autonomy of mental explanations with respect to physical explanations. Although Putnam's example of the square peg and the round hole helps us to understand the autonomy of macrophysical explanations with respect to microphysical explanations, for example, it does not seem to help us understand the autonomy of mental explanations with respect to physical explanations. But what motivates this argument is the much deeper point that mental states are variably realised, and this in itself does seem to present a strong case for denying that physicalism implies reductionism or explanatory exhaustiveness.

2.1b.The Irreducibility Of The Special Sciences

Fodor (1981) exploits this idea to argue that physicalism is only inconsistent with mental autonomy if we suppose that commitment to the primacy of the physical requires commitment to sets of reducing laws that link mentalistic explanation to physical explanation. The antecedent of these laws would contain a physical predicate and the consequent would contain a mental predicate. But if mentalistic explanations were reducible to physicalistic explanations, these laws would have to contain predicates which figure independently in the laws of the theory to be reduced and predicates which figure independently in the laws of the reducing theory; and being laws, they would have to state that the predicate of the reduced theory is nomologically coextensive with the predicate of the reducing theory. So if physicalism implies the reduction of the mental to the physical, then for every mental predicate there must be a coextensive physical predicate, and the generalisation which expresses this coextension would have to be a law. This would make it possible to deductively derive any given mental explanation from a physical explanation when it is combined with the relevant reducing laws.

Fodor's point is that the variable realisability of the mental implies that mental predicates may be coextensive with a disjunction of heterogeneous physical predicates, and this in turn implies that the mental cannot be explained in terms of the physical. It is extremely unlikely that the disjunction of physical predicates, with which any given mental predicate may be coextensive, can figure in an independent physical law to begin with, never mind as the antecedent of a reducing law, since the disjunction of physical predicates is unlikely to form a natural physical kind. Fodor illustrates the basic idea with the example of economics, a special science whose laws are irreducible to the laws of physics. Although it is clearly the case that every event which is a monetary exchange has a physical description, the physical description under which every one of these events falls must be wildly disjunctive, such that the likelihood of that description's figuring as the antecedent or consequent of a law of physics is pretty slim. The reason for this is that the set of events which count as a monetary exchange is classified as such for specific social and economic purposes, but this classification groups together events whose physical descriptions fail to display the unity of a natural physical kind. The upshot of all of this is that although every monetary exchange is identical with some physical event, it does not follow that economics thereby loses its

autonomy, and by the same token, although every mental event is identical with some physical event, it does not follow that the mental thereby loses its autonomy.

2.2a.The Problem With Realisation

Does the Putnam-Fodor line successfully secure the autonomy of the mental within the physicalistic framework? The fact that mental states can be realised in a wide variety of physical states and structures which, when taken together, fail to display the unity of a natural physical kind, does seem to decrease our chances of finding the type of bridge laws thought to be required to reduce the mental to the physical. But perhaps we ought to probe somewhat deeper at this point into the idea that mental states can be realised in physical states. For it is not so clear that we can say, as does Putnam, that the mental state of thinking about next summer's vacation can have a physical and chemical realisation in the brain. Again, the problem is that it is not obvious how a physical and chemical state of an individual's brain, whatever it happens to be, can be a realisation of his thinking about next summer's vacation, when it is intrinsic to his thinking that it is about a particular situation, whereas the physical and chemical state of the brain cannot be about anything. It might help if we have a more precise definition of realisation to work with. Here is Poland's:

> all attributes must be realised by physical attributes in the sense that configurations of physical objects and attributes constitute the instantiations of all attributes that are, or can be, instantiated in nature. (1994: 18).

Poland builds into this definition of realisation the requirement that the physical attributes that constitute the instantiation of mental attributes are nomologically sufficient for them. So what we are being asked to suppose here is that a particular configuration of physical properties is nomologically sufficient for it to be the case that the individual is thinking about next summer's vacation, just as a particular configuration of physical properties is nomologically sufficient for glass to be transparent or for water to be liquid. To be fair to Putnam and Fodor, this definition of realisation is perhaps too strong to capture their understanding of realisation, since they are committed to the gross unlikelihood of there being nomologically coextensive mental and physical properties. But it is not very clear that we can make sense of realisation without being able to say that certain configurations of physical properties are nomologically sufficient for the instantiation of certain mental properties, since nomological sufficiency does seem to be required to articulate the idea that mental properties are instantiated *in virtue* of the instantiation of physical properties. We

could, of course, simply assert that a particular configuration of physical properties is materially sufficient on a particular occasion, without thereby implying the satisfaction of laws. But regardless of whether we incorporate nomologicality into the definition of realisation, it does seem that if a configuration of physical properties realises a mental property, then that particular configuration must at least be materially sufficient on that occasion for the instantiation of the mental property.

Once again, however, the problem is that the intentionality intrinsic to the individual's thinking does not seem to be realisable in any sort of physical system, as liquidity and transparency clearly are. It is not too difficult to understand what it means to say that liquidity is realised in the molecular structure of water, or that transparency is realised in the molecular structure of glass, in the sense that they are nothing over and above these structures. But as I have already noted, such comparisons must fail to shed light on what it means to say that an individual's thinking can be realised in the physical or chemical states of his brain. It seems to me that realisation is a relation peculiar to the physical sciences, which applies straight-forwardly to physical phenomena. There seems to be no problem at all with saying that a given physical configuration is sufficient on a particular occasion for the instantiation of a given physical property, but I think it is problematic to extend this notion to cover mental properties, since it does not seem possible for any given physical configuration to constitute an individual's thinking about next summer's vacation. So it seems to me that the use of the notion of realisation to articulate the nature of the relation between the mental and the physical is questionable, even if it does appear to serve the purpose of blocking the implication from physicalism to the explanatory exhaustiveness of the physical sciences.

2.2b. The Possibility Of Local Reductions

Even if we grant the use of the notion of realisation for the sake of assessing the strength of the argument, in so far as it is construed as an argument to block this implication, there are still further problems which come to light. One central problem is that far from providing the physicalist with a means of avoiding the reductionist implications inherent in the assumption that everything which exists is physical, the variable realisability of the mental can be made out to be consistent with the reduction of the mental to the physical, if the reductions are restricted to specific species. That physicalism does not entail the explanatory exhaustiveness of the physical sciences appears to be blocked by the point that mental states can be variably

realised across different species (mammals, reptiles, extra-terrestrials, etc.), but it is arguably the case that this is consistent with there being a number of different species-specific reductions, weakening the overall anti-reductionist implications of the claim that mental states can be variably realised.

Kim (1993a: 273) points out that if the notion of realisation is to be cashed out in terms of conditionals of the form $p \Rightarrow m$, which are species-specific, then it must be possible to generate bridge laws which can be used to effect the local reduction of mental states to physical states. What this amounts to is that if we assume that the realisation of the mental by the physical entails that conditionals of the form $p \Rightarrow m$ hold (which means that (i) if x is in physical state p, then x is in mental state m, and (ii) the physical state p which realises the mental state m is nomologically sufficient for it), and we restrict these conditionals to specific species, then it looks as if we can say that, within any given species, s, p is both necessary and sufficient for m. This will then provide us with the restricted bridge law, $s \Rightarrow (p \Leftrightarrow m)$. If we follow this tactic, it seems that local reducibility can be made out to be consistent with the variable realisability of the mental, diminishing the strength of its over-all anti-reductionist implications.

If we take it for granted, for the sake of the argument, that the notion of realisation is unproblematic, and if it is plausible to make the move that Kim suggests, then we seem to have secured the possibility of generating species-specific bridge laws which may be used to carry out local reductions of the mental to the physical. It seems that the attempt to secure the autonomy of the mental by appealing to the possibility of variable realisation simply serves to highlight the difficulty with using the notion of realisation as a way of articulating the nature of the relation between the mental and the physical. However, it might be replied that this move is not necessarily going to work, since it might be the case that mental states are variably realised *within* specific species. Given the phenomena of maturation and development, injuries to the brain, and so on, it is not so clear that species-specific reductions are going to be available. Individuals suffering damage to certain parts of the brain may thereby suffer loss of memory, for instance, which may be regained when other parts of the brain begin to compensate for this damage by fulfilling a role previously fulfilled by the damaged parts. Given cases of this type, and perhaps many others, there seems to be no reason to insist that individuals within a specific species who happen to be in the same neurophysiological state, p, will necessarily be in the same mental state, m. But as Kim

correctly points out, even once these differences are taken into account, and even if they are more wide-spread than it is realistic to assume, in order to make use of the notion of realisation as a way of articulating the relation between the mental and the physical, it must be assumed that the psychology of each of us is at least *locally* reducible to his *own* individual neurobiology. Kim writes that:

> the conclusion we must draw is that the multiple realisability of the mental has no antireductionist implications of great significance; on the contrary, it entails, or at least is consistent with, the local reducibility of psychology, local relative to species or physical structure-types. If psychological states are multiply realised, that only means that we shall have multiple local reductions of psychology. The multiple realisation argument, if it works, shows that a global reduction is not in the offing; however, local reductions are reduction enough, by any reasonable standards and in their philosophical implications (1993a: 275).

The problem lies with the metaphysical implications of using the notion of realisation to articulate the nature of the relation between the mental and the physical. For although the notion of realisation does seem to have anti-reductionist implications, in the sense that it is consistent with the variable realisability of the mental, this cannot be sufficient to secure the autonomy of the mental, since the realisation of the mental by the physical is also consistent with the possibility of making local reductions of the type envisaged above. There does not seem to be anything to rule this out. Admittedly, however, the argument as Kim presents it cannot be conclusive, since it does not yet show that the variable realisability of the mental *entails* local reducibility, which would be required to refute the anti-reductionist stand. Yet, even if it is admitted that local reductions are not entailed by the variable realisability of the mental, they do seem to be consistent with it; once this is acknowledged, the argument to block the implication from physicalism to reductionism, if not refuted, must certainly be weakened somewhat. And the problem, I suspect, lies with using the notion of realisation to articulate the nature of the relationship between the mental and the physical.

2.3.Psychophysical Anomalism

Perhaps a more plausible argument to block the implication from physicalism to reductionism is suggested by Davidson (1980a). Davidson's central claim is that mental properties cannot be reduced to physical properties because there cannot be a strict law-like statement that correlates these properties. This is worth considering, starting with the following question: what is it about mental properties that prohibits their figuring in nomological relations with physical properties, in such a manner that would permit the

generation of bridge laws between the mental and the physical? The central idea is that mental properties of events are identified as the mental properties they are by the patterns of rationalistic and justificatory relations that give structure to the logical and epistemic space in which they are instantiated. Physical properties, on the other hand, are identified as the physical properties they are by the patterns of causal relations that give structure to the domain which is subsumed by the closed and deterministic theories of physical science. So the reason why statements that link mental and physical properties cannot be strict laws is that it would thereby be possible to infer the instantiation of mental properties on the basis of the instantiation of physical properties, which would effectively mean that it would be possible to identify mental properties without attention to the rationalistic and justificatory relations which in fact identify them. As Davidson puts it:

> There are no strict psychophysical laws because of the disparate commitments of the mental and the physical schemes. It is a feature of physical reality that physical change can be explained by laws that connect it with other changes and conditions physically described. It is a feature of the mental that the attribution of mental phenomena must be responsible to the background of reasons, beliefs and intentions of the individual. There cannot be tight connections between the realms if each is to retain allegiance to its proper source of evidence. (1980a: 222).

What lies in the background here is a very plausible assumption concerning the attribution of intentional states and events to individuals: their attribution is necessarily governed by the constraints of rationality and normativity, such that an individual cannot be said to have the thought 'that the candle has blown out', for example, unless he is also capable of having various other logically related thoughts and beliefs, which identify this particular thought by locating it in a logical and epistemic space. The logical space which identifies this thought as having the content it has is structured by the interlocking patterns of rational and justificatory relations that hold between the contents of the rest of the individual's thoughts and beliefs. Therefore, we cannot attribute this particular thought to the individual without taking into account the possibility of his having the appropriately structured background. Furthermore, we cannot hope to reconstruct this pattern of relations at the level of physical description, and this point certainly seems to suggest a way of supporting the claim that the mental is autonomous with respect to the physical.

Davidson combines his arguments for the autonomy of the mental with a rather interesting, but questionable, argument for the token identity claim. Indeed, what is interesting about this argument is that it starts out with the assumption that the mental is autonomous, and it concludes that mental events *must* be identical with physical events. But why should the

autonomy of the mental be thought to imply the truth of the token-identity thesis? Why should the autonomy of the mental, when made out in terms of the impossibility of subsuming mental and physical properties under laws, have physicalistic implications?

The argument is simply that mental events enter into causal relations with physical events, as when the desire for a drink causes an individual to walk to the water fountain; but if events enter into causal relations, according to Davidson, there must be a description of these events under which they instantiate a law. This is required by his commitment to the nomological conception of causality, which states that causal relations between events must instantiate law governed regularities. The problem is that since the autonomy of the mental has been secured by the principle of psychophysical anomalism, which depends on the assumption that mental properties of events cannot figure in laws, it follows that causal relations in which mental events are involved must instantiate a physical law. The final step in the argument is to point out that if mental events are subsumable by physical laws, then they must have a physical description, in virtue of which the physical laws apply, and for that reason it follows that mental events must be physical events.

This conclusion appears to be necessary, given that the autonomy of the mental implies the failure of mental properties to figure in laws. But although the conclusion is that every mental event must be a physical event, this does not yet tell us how mental and physical properties are related. Davidson's suggestion is not that mental properties are realised in physical events, which I welcome, but simply that mental properties of events are supervenient on the physical properties of events:

> mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect...supervenience of this kind does not entail reducibility through law or definition. (1980a: 214).

2.4a. The Problem With Fusions

Davidson's argument is rather complex, but as I suggested, it is questionable at certain points. I want to begin with his argument for the identity claim, and I want to suggest that the identity claim can be dispensed with. The claim that mental events are identical with physical events can be seen to analyse into at least two components: first, some events have both mental and physical descriptions; second, physical descriptions of events are more basic than their mental

descriptions. The first general point I want to raise here is that it is not immediately clear what it means to say that physical events can have mental descriptions; or at least, it is not obvious how we are to understand the claim that the mental event of 'thinking that it is going to snow', for example, is identical with an event which has a more basic description in physical terms. The problem is that it is not clear how this description could be correctly or incorrectly applied to a physical event, unless there were some way of picking out the relevant physical event, and then applying this description on the basis of its physical properties. But this cannot be what Davidson has in mind, since he builds his argument on the assumption that there can be no such move. The description of the event in mental terms can only be justified in terms of its location in the larger network of mental events. However, this surely takes us to a deeper point, which is that there does not seem to be any justification for making the identity claim in the first place, other than adherence to the nomological conception of causality. This gives us a possible line of criticism, since the identity claim will have to be made out in such a way as to cohere with the nomological conception of causality.

Hornsby (1980-1) suggests that the only way to make sense of the identity claim, given that it is motivated by the nomological conception of causality, is to think of mental events as being mereologically composed out of physical events. The point is that the types of events which are subsumed under strict laws cannot be everyday macroscopic events, like thunderstorms and boating trips on the lake, but must rather be the microscopic events out of which these events are mereologically composed. If macroscopic events are to be related as cause to effect, then the nomological conception of causality demands that there be a description of these events under which they are subsumed by a strict law. So if the thunderstorm caused the boating trip to be cancelled, there must be a description of these events under which they are law instantiating. But the problem is that since macroscopic events of this type are too coarsely individuated to be subsumed by strict laws, the nomological conception of causality can only be supported if macroscopic events can be said to be composed out of 'fusions' of microscopic events, which *are* individuated finely enough to be subsumed by strict laws.

The same reasoning applies to mental events. If the mental event of thinking that it is going to snow caused the intentional act of putting on a warm coat, then there must be a description of these events under which they are subsumed by a strict law. But given that macroscopic events of this type are too coarsely individuated to be subsumed by strict law, they too must be said to be composed out of 'fusions' of microscopic events, which are individuated finely

enough to be subsumed by strict law. So if the autonomy of the mental implies token physicalism, on the strength of the nomological conception of causality, then it seems to follow that mental events are identical with physical events in the sense that they must be composed out of fusions of neural events.

Hornsby's (1980-1: 82-85) central point is that the mereological conception of events is incoherent, and that it therefore cannot be appealed to in order to support the nomological conception of causality. The problem with the mereological conception of events is that fails to set out any principles of construction, on the basis of which neural events can be combined to form the fusions with which mental events are to be identified. This has the unhappy consequence that a variety of *ad hoc* fusions could be formed simply by combining neural events in an arbitrary manner, possibly resulting in a combination of neural events, some of which having no obvious relevance to the mental event with which it is to be identified. More than this, the mereological conception of events does not seem to provide us with any principles on the basis of which we could distinguish those fusions which are genuine events from those fusions which are merely *ad hoc* constructions, and this is connected to the fact that the mereological conception of events fails to impose principled restrictions on the relations that the individual neural events must bear to each other, if they are to be combined to form the fusions with which mental events are to be identified. But presumably, such principled restrictions could not be provided without admitting at least the possibility of generating reducing laws of the type the anti-reductionist needs to avoid. So if adopting the mereological conception of events is the only way of supporting the nomological conception of causality, then we have good reason for giving up that conception of causality; and once that is given up, we have immediately lost one central means of demonstrating the consistency of mental autonomy with physicalism. That is, once the nomological conception of causality is deemed insupportable, we have to abandon this particular route from the autonomy of the mental to the token identity thesis.

2.4b.Defending Non-Reductivism: Property Coinstantiation

Is this argument too quick? Macdonald (1989) thinks it is. She points out that even if the mereological conception of events cannot be made plausible, this need not force us to abandon the principle of the nomological conception of causality, and if we are not forced to abandon that, then it is simply not the case that we have undermined this argument for non-

reductive physicalism. Macdonald's objection is that it is a mistake to say that the nomological conception of causality requires that the relation between mental and physical events be one of mereological composition; there is an alternative way of conceiving the relation between these events which explains how mental events can be covered by strict laws:

> laws relate events in terms of their properties. In particular, they relate events of certain types in virtue of those events instancing properties of certain types. We may have good reasons for thinking that event types like the type 'being an avalanche' have associated with them certain properties distinct from any of the properties associated with event types which their so-called 'parts' instance, i.e., that the properties associated with macrophysical and microphysical event types are distinct. But this only leads to the conclusion that tokens of macrophysical event types are not covered by laws if one assumes that the instancings by events of those very properties which make for the distinctness of avalanches from their 'parts' (e.g., the property of being an avalanche) are themselves distinct from instancings of the properties associated with such events' parts which do figure in laws. (1989: 173).

This is a very dense passage, but let me try to paraphrase it somewhat. Macdonald seems to be arguing that events are covered by laws in virtue of instantiating microscopic property types. Macroscopic property types are distinct from microscopic property types, and this makes it look as if macroscopic events are not subsumable under law unless they are mereologically composed out of fusions of microscopic events. But by Macdonald's lights, there is no need to assume that this is the case, since it can be argued that the instantiation of microscopic properties, by virtue of which events are subsumed under law, is at the same time the instantiation of macroscopic properties. It is only if the instantiation of tokens of microscopic properties is distinct from the instantiation of tokens of macroscopic properties that the application of the nomological conception of causality to macroscopic events will be problematic. If these properties are coinstantiated, then tokens of macroscopic events will not be distinct from tokens of microscopic events; and hence it will follow that laws can apply to tokens of macroscopic events in virtue of the fact that they stand to tokens of microscopic events in the relation of property coinstantiation.

So Macdonald's suggestion is that the relation between mental and physical events, in virtue of which mental events can be covered by laws, is one of property coinstantiation. The crucial point is that since the question of whether the relation of property coinstantiation holds is silent on the issue of *how many* microscopic events may constitute a given macroscopic event, the nomological conception of causality can be supported independently of assuming the mereological conception of events. Before I question the intelligibility of defending the mental-physical identity claim with the claim that macrophysical and

microphysical properties can be coinstantiated, let me consider one of Macdonald's more recent attempts to articulate this analogy more precisely.

2.4c.The Defence Deepened: An Analogy From Biology

Macdonald (1995) deepens her defence of non-reductive physicalism by developing a general anti-reductionist strategy which bears some resemblance to the strategy adopted by Fodor. It involves showing how the irreducibility of properties which figure in the laws of special sciences to properties which figure in the laws of physics can serve as an analogy for the irreducibility of mental properties to physical properties. Specifically, Macdonald's tactic is to explain the irreducibility of mental properties to the physical properties that instance them by analogy with the irreducibility of biological properties to the physicochemical properties that instance them. The important point is that biological properties are identified by a pattern of relations that is distinctively different from the pattern of relations that identify physicochemical properties, even though an instance of a biological property is in fact claimed to be an instance of some physicochemical property. Biological properties are identified by their functional role, whereas physicochemical properties are identified by their causal-nomological role. Biological properties are therefore dependent on physicochemical properties in the sense of being coinstantiated by them, yet are irreducible to them in the sense of having a nature which cannot be exhaustively replicated at the level of physicochemical description.

But if this is to serve as a way of understanding the mental-physical relation, a more precise specification of the relation between biological and physicochemical properties is required, one which clarifies what it means to say that macrophysical properties can be instanced by microphysical properties. This will also be relevant to Macdonald's argument to avoid fusions. She offers the following example. Suppose that three organisms all have bottle-green colouring, a chameleon, a butterfly, and a bird. All three can be described from a physicochemical point of view in the very same terms, but they cannot be described from a biological point of view in the very same terms. The chameleon's bottle-green colouring serves as camouflage. The butterfly's serves as a warning to predators that it is more or less inedible. The bird has no biological description in virtue of its bottle-green colouring at all. So the pattern of causal relations between the physicochemical properties that underwrite the bottle-green colouring possessed by these organisms, in Macdonald's view, is undisturbed by

the different functional patterns produced at the biological level; in this respect, biological properties have a specific nature which cannot be exhaustively characterised by an exhaustive characterisation of the physicochemical properties which instance them. But what is the relationship that obtains between biological and physicochemical properties and their instances, which helps us understand the relation between mental and physical properties and their instances? Macdonald writes:

> Given that biological properties arise as the result of natural selection operating on these instances of physicochemical properties, and given that physicochemical properties acquire biological functions as a result of the process of such selection, the most plausible account of the relationship is that to instance a biological property, say the property of having aposematic colouring, just is to instance the property of being bottle-green in colour, given that the latter instance has the causal history it does. It seems, in short, that instances of biological properties just are instances of certain physicochemical properties. (1995: 149).

It seems to me that this does not really further our understanding of what it means to say that mental properties can be coinstantiated by physical properties, since the argument appeals to a relation which is restricted to the macrophysical domain, as opposed to a relation between the macrophysical and the microphysical. It cannot explain the relation between microphysical properties and macrophysical properties to say that the biological property of having aposematic colouring is instanced by the property of being bottle-green. Does this not remain within the macrophysical domain? To clarify the relation between the macrophysical and the microphysical, or between the mental and the physical in particular, the property of having aposematic colouring would have to be shown to be instanced by some of the physicochemical properties at the microphysical level. Otherwise, the irreducibility of biological properties to the physicochemical properties that instance them has not explained the irreducibility of mental properties to the physical properties that instance them.

The analogy also breaks down in other ways. Although it is fairly clear what it means to say that physicochemical properties acquire biological functions as the result of the process of natural selection, it is not so clear which processes are supposed to be responsible for the fact that physical properties acquire their mental descriptions. Nor is it very clear that mental properties can be said to arise as the result of the operation of these processes, whatever they are, on instances of physical properties, in the same way that biological properties can be said to arise as the result of natural selection operating on instances of physicochemical properties. But unless these questions can be answered, it seems to me that the claim that macrophysical and microphysical properties can be coinstantiated does not further our understanding of

what it means to say that mental and physical properties can be coinstantiated. For these reasons, I think that Macdonald fails to articulate the relation between the mental and the physical in such a way that avoids Hornsby's central objection to the identity claim.

3.Supervenience

3.1.Weak And Strong Definitions

The notion of supervenience has more recently been appealed to in order to give expression to the relation between the mental and the physical in a manner deemed satisfactory to the non-reductivist. The reason for this is that supervenience is thought to embody the general idea that the mental is dependent on the physical without being reducible to it. The fact that one set of properties supervenes on another set, in this case mental properties on physical properties, is thus expressed in the following way: no two events can differ with respect to their mental properties without differing with respect to their physical properties, such that any difference with respect to mental properties will be accompanied by some difference with respect to physical properties. Expressed in this way, the notion of supervenience seems to be a rather *weak* statement of the relation between the mental and the physical, since what it amounts to is that mental properties and physical properties stand in a relation of covariance. Its weakness is what makes it attractive to the non-reductivist. The mere fact that mental properties covary with physical properties is not sufficient to warrant the introduction of psycho-physical laws to link them, and this is precisely what the non-reductivist is looking for. However, its weakness is also a problem: it does not seem to offer an adequate statement of physicalism. A mere statement of psycho-physical property covariance is too weak to capture the idea that the mental is *dependent* on the physical, since it does nothing more than report occurrent patterns of change between different sets of properties.

Kim (1993b: 148) points out that the notion of dependence is in fact an additional component of supervenience, which takes us beyond the mere statement of property covariance. The statement of property covariance reports patterns of change between sets or families of properties; but this does not yet answer the deeper question, which is why these families of properties covary as they do. The problem is that covariance itself does not entail dependence, since the mere report that two families of properties stand in the relation of covariance does not entail that the variation in one set is dependent on variations in the other,

let alone that one family is more basic with respect to the other. As it stands, the primacy of the physical, and the dependence of the mental on the physical, cannot be adequately expressed by the statement of mere covariance, unless that statement is boosted by an explanation of why such covariance holds. But boosting the statement in this manner must make it difficult to appreciate the sense in which physicalism is consistent with the autonomy of the mental. Arguably, a statement of dependence involves a modal claim, which gives rise to the possibility of generating laws of the type required to reduce the mental to the physical.

Kim's distinction between weak and strong covariance might help to illustrate the problem. Weak covariance states that, necessarily, if anything has property F in A, there exists a property G in B such that the thing has G, and everything that has G has F. Strong covariance states that, necessarily, if anything has property F in A, there exists a property G in B such that the thing has G, and necessarily, everything with G has F. Thus, in order to count as a dependence relation, weak covariance has to at least be strengthened into strong covariance by the incorporation of the modal operator 'necessarily' into the final clause of its definition. In terms of mental and physical properties, weak covariance states that, necessarily, if an event has a mental property m, there exists a physical property p such that the event has p, and every event that has p has m. Strong covariance states that, necessarily, if an event has a mental property m, there exists a physical property p such that the event has p, and necessarily, every event with p has m.

The result of this is that strong covariance guarantees that the correlation of mental and physical properties holds stable across possible worlds, whereas weak covariance restricts its guarantees to the given world under consideration. So the central problem with weak covariance is that it restricts its constraints on the distribution of mental properties only within any given world, and as such it fails to capture the idea that the physical facts determine all the facts, since that would seem to require that the constraints on the distribution of mental properties held across every possible world. Furthermore, since weak covariance requires only accidental connections between mental and physical properties, it does not seem sufficient to express the idea that an individual has his mental properties in virtue of his physical properties. Poland complains that such relations:

> underwrite no counter-factual truths regarding what non-physical properties an individual would have if he were to have certain physical properties...There is no required isolation of all relevant physical attributes that are nomologically sufficient for the realisation of given non-physical attributes...Thus a weak supervenience approach to the formulation of physicalism fails to be adequate. (1994: 81-2).

Weak supervenience thus appears to be an inadequate expression of physicalism. The implication is that in order to count as the expression of a dependence relation, we have to strengthen the definition of supervenience by strengthening the relation of weak covariance to yield the relation of strong covariance. This would at least provide a guarantee that the psycho-physical property correlations held across worlds, such that the instantiation of a given set of physical properties ensured the instantiation of a given set of mental properties. The problem with this, however, is that the required strengthening of the covariance relation seems to yield psycho-physical laws of the type the non-reductionist must be concerned about. That is, it seems that the only way of giving adequate expression to the physicalist's dependence principles at the same time gives expression to the reductionist implications of these principles. This will clearly be welcomed by reductionists, since it seems to threaten the case for saying that physicalism is consistent with the autonomy of the mental. A key point in the case for anti-reductionism is the impossibility of stating strict psycho-physical laws that would allow the mental to be nomologically reduced to the physical. But once it is claimed that physicalism cannot be adequately stated unless the instantiation of specific physical properties guarantees the instantiation of specific mental properties across worlds, it begins to look as if physicalism is in fact inconsistent with mental autonomy.[3]

3.2.A Global Definition

One way of responding to this threat is to define supervenience in more global terms, so that a statement of the dependence of the mental on the physical is to be assessed for adequacy by comparing whole worlds as opposed to individuals within worlds. The suggestion is that the

---

[3] Admittedly, however, reductionism might not immediately follow from the use of strong supervenience as an expression of the relation between the mental and the physical. McLaughlin (1995: 47) points out that there are some cases involving nomologically coextensive properties that would not be counted as cases of reduction. Electrical conductivity properties of metals are nomologically necessary and sufficient for their thermal conductivity properties, since the same free electrons carry charge and heat, yet the thermal conductivity properties of metals do not reduce to their electrical conductivity properties. The reductionist implications are indeed there to be drawn out, but whether they present a refutation of the combination of mental autonomy with physicalism has not been conclusively established. However, compare this with Grimes (1995: 112-3), who insists that the set of one-way conditionals, $G \Rightarrow F$, for instance, which arise out of the definition of strong supervenience, are enough for reduction, since they allow us to derive every instantiation of every supervenient property on the basis of the instantiation of some subvenient property. On this view, which trades on the fact that there is no one form of reduction, and that one form may be stronger or weaker than others, necessary coextension is not even required. Grimes adds that, even if this stronger definition fails to imply a relation of necessary coextension between mental and physical properties, it still seems too reductive to serve as a general vehicle for advancing the nonreductivist agenda.

mental supervenes globally on the physical in the sense that any two worlds which are indiscernible with respect to the distribution of physical properties are indiscernible with respect to the distribution of mental properties. The advantage of defining the dependence of the mental on the physical globally is that it is silent on the strength of specific mental-physical property correlations, the important consequence of which is that global supervenience does not seem to threaten reductionism in the way that strong supervenience seems to. As a means of avoiding the reductionist implications inherent in strong supervenience, therefore, this definition must be more effective. Whereas strong supervenience requires that any two individuals who are indiscernible with respect to their physical properties must be indiscernible with respect to their mental properties, global supervenience is perfectly consistent with the failure of such indiscernibility, and hence with the failure of generating appropriate sets of reducing laws, since its central intuition is that mental properties are determined in a holistic, rather than an individualistic, manner.

However, as an expression of physicalism, global supervenience seems rather inadequate. The problem is simply that it does not seem strong enough to express a satisfactory dependence relation between the mental and the physical. Whilst its broadness is certainly to its advantage, in so far as it precludes the possibility of generating sets of reducing laws covering specific mental and physical properties, it is also to its disadvantage, in so far as this unspecificity makes it too permissive to qualify as a useful relation of dependence. The problem, similar to that encountered by weak supervenience, is that it fails to be specific concerning which physical properties are relevant to the determination of which mental properties. Global supervenience thus seems to be perfectly consistent with there being two physically indiscernible individuals within any given world who have radically divergent mental properties, or even one with and one without any mental properties, since what matters is only that these individuals could not be inhabitants of two physically indiscernible worlds. Furthermore, it is often claimed that the global supervenience of the mental on the physical is perfectly consistent with there being two worlds which differ in the minutest physical respect, but where two physically indiscernible individuals, one from each world, are drastically different in mental respects. Smith, for example, puts the complaint thus:

> since it fails to put any restrictions on the psychological properties that are instantiated in worlds that are not physically identical, it is compatible with there being a situation differing physically from the actual one only in the existence of an extra penguin, but where the psychological facts are as different as you like from the actual ones. We need, somehow, to narrow down to the *relevant* base properties that determine the supervenient ones. (1993: 238-9).

The force of this objection seems to be that since global supervenience fails to exclude those physical facts which are irrelevant to the determination of psychological facts, it is too permissive to qualify as a dependence relation of the mental on the physical. The onus is therefore placed on the physicalist to find a means of excluding those physical facts which are completely irrelevant to the determination of an individual's psychological facts, such as the existence of an extra penguin in the antarctic, or the existence of one extra hydrogen atom somewhere in deep space, without which global supervenience would seem to lose its credibility as a statement of the dependence of the mental on the physical.

According to Post (1995), this objection is not as strong as it seems. What the physicalist needs to exclude as irrelevant to the physical determination of psychological facts are only those physical facts which do no causal work in this respect. This means that he needs to restrict the physical determination base, whilst remaining committed to the global determination of the mental by the physical, so as to exclude such trivial differences from being relevant to the determination of an individual's psychology. Here is how Post thinks this can be achieved. Suppose that T is the set of physical conditions that determines whether certain psychological facts are true of a given individual, x. Then the physical fact, that there exists one extra hydrogen atom in x's world, let us call it Φ, can be regarded as irrelevant to the determination of x's psychology if this fact could have been excluded from T without disrupting the determination at work in this case. Post points out that if there is a proper subset Δ of T, which does not contain condition Φ, but which nevertheless determines whether certain psychological facts are true of x, then we can say that condition Φ is irrelevant to the fact that x has the psychology he has. Condition Φ is therefore relevant to the determination of x's psychology only if it is a member of the *least set* Δ of T, which is uniquely sufficient on this occasion to determine whether certain psychological facts are true of x or not.

The upshot of this is that when x's world is compared against a physically indiscernible world, but for the fact that it differs only in the existence of one extra hydrogen atom somewhere in deep space, we should not be forced to admit the possibility of radical psychological differences for x according as he is considered in one possible world or the other, because condition Φ can be excluded from the least set Δ, as doing no real work in this respect. Given the existence of a possible world, therefore, which differs minutely from this

one, if the conditions which belong to the least set Δ obtain in both, then the psychological facts which are true of x in this world are true of him in that world. And furthermore, despite the fact that the physical conditions have been narrowed to exclude irrelevant factors, this need not be thought to affect the *global* nature of the dependence relation. For as Post puts it:

> Provided there being an extra hydrogen atom somewhere in deep space is not a member of the least set to determine whether I have fruit fly mentality, it is irrelevant to the determination. And, of course, Φ can be a member of such a least set, hence relevant, whether or not Φ is a property or relation of the individual x, and whether or not Φ is minute, synchronous with x, in spatiotemporal proximity to x, or we require that whole worlds never be compared for in-/discernibility in regard to sets of properties. (1995: 93-94).

It might raise some concern that although this move does seem to be effective in excluding irrelevant factors from figuring in the physical determination of the mental, it does not yet provide us with a means of isolating the least sets of relevant conditions which are claimed to be important. The reason for this is that it still does not provide us with a means of isolating the specific physical conditions which are relevant to the determination of specific mental conditions. The unspecificity of the global definition of supervenience leaves us without a useful method of isolating the physical conditions which are irrelevant, since it leaves us without an understanding of which mental conditions obtain in virtue of which physical conditions. But is it too quick to think that this omission is necessarily a weakness? Post believes that it is. By Post's lights, the central point of the global definition of the physical determination of the mental is only to provide us with a general or sweeping statement of the relation between the mental and the physical. It is not one of its jobs to provide us with a statement of how the physical determines the mental in every case, or of how specific physical conditions determine specific mental conditions. In other words, to complain that global supervenience is too permissive because it is unspecific and unfocused, in the sense that it does not provide us with a means of isolating the specific physical conditions which are relevant to the determination of specific mental conditions in any given case, would be to miss the point.[4]

---

[4] It is not obvious that this objection does miss the point. If physicalism is to offer a way of understanding how the mental is itself part of the physical world, rather than simply being connected to it by way of natural relations, then it is not unreasonable to insist that a stronger claim has to be made. With respect to this point, Witmer (1998: 85) states that: "However exactly we are to understand physicalism, it is clear that physicalism is meant to be a claim about the metaphysical character of such phenomena as mind and society; that is, it is meant to be a claim about what such things are...What makes physicalism the tough-minded claim that it purports to be is that it goes beyond the banal and obvious claim that mind and society and so forth are linked to the physical by natural relations of some sort...Physicalism is, further, a claim about the nature of the items so linked to the physical."

Whether this objection misses the point or not, and I am inclined to think it does not, there is still a further difficulty which global supervenience faces. The difficulty is that even if we had this method of exclusion available to us, and even if we were in the position to implement it successfully, global supervenience is open to the objection that it does not seem to capture the physicalist's central intuition that the physical facts determine all the facts. This can be illustrated if we imagine a case in which there is a possible world whose physical indiscernibility from this one is complete in every respect, but where there is nonetheless a significant divergence at the level of the psychological. Such divergence should not be permitted at all under global supervenience if it is to count as a plausible statement of the physicalist's claim that the physical facts determine all the facts.

Moser and Trout (1995) imagine a case in which there is psychological divergence across physically indiscernible worlds, which is effected by the fact that in one of these worlds there are atypical psychological laws. Suppose that there is a world which is distinctive in that its laws of psychological causation, for example, give rise to a level of psychological fact which is at odds with the level of psychological fact characteristic of every other physically indiscernible world. The physical facts in each world are indiscernible, and as such each determines that the same first-level psychological facts obtain. But given that there are different laws of psychological causation operative in this world, the first level psychological facts then determine a second level of psychological facts which are divergent with respect to the second level psychological facts determined by the first level facts in those worlds whose laws of psychological causation are not atypical. In this world, psychological events of assenting, for example, do not generate dispostional or habit-like belief-states and intention-states, and for this reason the imagined situation seems to display psychological uniformity at the first level, but psychological divergence at the second, despite the fact that the worlds under consideration, and hence the individuals within these worlds, are physically indiscernible in every respect. Moser and Trout claim that:

> We have no compelling reason to hold that in *all* physically possible worlds with the same physical conditions and physical laws, the same laws of psychological causation always obtain. It seems quite plausible to suppose, after all, that the aforementioned case of psychological divergence at a secondary level involves physically possible worlds. (1995: 199).

The possibility of psychological divergence in physically indiscernible worlds does seem to drastically undermine the chances of giving a statement of the relation between the mental and the physical which respects the principle that the physical facts determine all the facts,

whilst at the same time respecting the autonomous nature of the mental. It does not seem plausible at all to insist that the physical facts determine all the facts when it is possible that distinct sets of psychological laws could be operative in two physically indiscernible worlds, determining a secondary level of psychological fact in one world which is not replicated in the other. So whereas global supervenience seems to escape the charge that it is too permissive because it allows radical psychological divergence in virtue of some minute physical divergence, it does not seem equipped to deal with the objection that there could be radical psychological divergence even though there is no physical divergence *whatsoever*.

Indeed, to make matters worse, such a case suggests that it does not seem possible to retain the central physicalist principle at all, since it is conceivable that there are psychological facts which are not in any sense determined by underlying physical facts. It would perhaps require some effort to fill in the details of this scenario, but in so far as we have no contrary evidence to suggest that such a world is not physically possible, I think we have reasonable grounds for doubting the truth of the claim that the physical facts determine all the facts. It seems to me, therefore, that there are some genuine difficulties with achieving the required balance between physicalism and autonomy, which the physicalist must be able to deal with in a convincing manner, in order to explain how his claim that the physical facts determine all the facts does not undermine the claim that the mental is autonomous with respect to the physical.

4.Conclusion

In its attempt to secure the autonomous nature of mentalistic explanation, non-reductive physicalism places a great deal of weight on the issue of the irreducibility of mental properties to physical properties. Its statement of the relation between the mental and the physical draws on the metaphysical principles which seem to guarantee the autonomous nature of mentalistic explanation by guaranteeing the irreducible status of mental properties to physical properties. Or in other words, the irreducibility of the mental to the physical is secured by the imposition of an order onto the world that purports to be consistent with the instantiation of irreducible mental properties in physical events. I have been arguing that this position is problematic. I have tried to demonstrate certain weaknesses in non-reductive physicalism by testing out some of the key arguments designed to show that the existence of an autonomous mental domain can be consistent with the embeddedness of the mental in the

physical structure of the world. To my mind, the problem lies with the non-reductivist's statement of the relation between the mental and the physical, which is dictated by his commitment to the metaphysical principles of physicalism. In attempting to articulate this relation, the non-reductivist faces the threat of reductionism. It is perhaps for this reason that the latter is keen to emphasise that he is not required to give a precise statement of how the mental is embedded in the physical structure of the world, but it seems to me that this is tantamount to admitting that the mental cannot be embedded in the physical structure of the world.

Yet even if we allow the non-reductivist the security he claims, there is a further problem which I believe he ought to take more seriously. Non-reductive physicalism is certainly sensitive to the fact that a satisfactory statement of the relation between the mental and the physical will be one which recognises the constraint presented by the autonomous nature of mentalistic explanation. However, I want to argue that there is a deeper sense in which mentalistic explanations are autonomous than he is in the position to acknowledge. Mentalistic explanations are autonomous in the deeper sense that they can be said to carry out the work required of them without implicating the explanatory resources of the physical sciences. To appreciate the deeper sense of the autonomous nature of mentalistic explanation it is therefore necessary to find a way of maintaining the complete separateness of the work carried out by mentalistic explanation from the work carried out by physicalistic explanation. But that cannot be achieved, nor even considered a possibility, in so far as there is a commitment to the metaphysical ordering of reality inherent in the non-reductivist's framework.

The metaphysical ordering of reality is integral to the physicalist's commitment to the claims that the physical facts determine all the facts, and that everything that occurs does so in virtue of what is occurring in the physical domain. To the physicalist, this must place a more basic constraint on what should count as a satisfactory statement of the relation between the mental and the physical. He can recognise the autonomous nature of mentalistic explanation only to the extent that it is consistent with his metaphysical ordering of reality. But this means that, in articulating the relation between the mental and the physical, he must give priority to his physicalistic commitments first and foremost, and the constraint presented by the autonomous nature of mentalistic explanations, to which he is certainly sensitive, then has to be worked around this starting point. The problem with this, however, is that it does not do

justice to the full extent of the autonomous nature of mentalistic explanation, since the work carried out by mentalistic explanations will always have to be underwritten by the work carried out by physicalistic explanations.

Suppose we want to explain why an individual is felling a tree. We might do so by saying that he thinks the tree would provide sufficient fire wood for his camping holiday. But to the physicalist, the explanation of his action at this level cannot be a complete explanation, since that would mean that the truth of this mentalistic explanation were fully independent of the truth of some physicalistic explanations, and it would also mean that the mentalistic explanation did not stand in need of support from such underlying physicalistic explanations. But this cannot be tolerable to the physicalist, since it would interfere with his principles that the physical facts determine all the facts, and that there is a complete and deterministic explanation of every occurrence in terms of occurrences at the level of micro-physical processes.

The crux of the matter for the physicalist is that the mentalistic explanation picks out only certain aspects of the individual's behaviour, which it does without being able to explain how that behaviour occurred. So given that the mentalistic explanation only seems to offer an alternative means of describing what the more basic physicalistic explanations already completely account for, it follows that the mentalistic explanation is not yet sufficient to fully explain the individual's action of felling the tree, regardless of the fact that it serves to illuminate an ineliminable feature of it. In order to complete the explanation in the right way, it would be necessary to show that the mentalistic explanation converges on the same subject matter as the physicalistic explanation of the occurrence of the individual's behaviour. So the immediate consequence of retaining the commitment to physicalism, whilst trying to recognise mental autonomy, is that the work carried out by mentalistic explanations turns out to be parasitic on the work carried out by physicalistic explanations. But even if this succeeds in recognising a limited degree of mental autonomy, we will have failed to recognise the deeper extent of the autonomous nature of mentalistic explanations. And if I am correct, non-reductive physicalism must therefore fail to yield a satisfactory statement of the relation between the mental and the physical. What is required to recognise the deeper extent of mental autonomy, how this might be achieved, and what it implies about the nature of the relation between the mental and the physical, will be topics for the remainder of this thesis.

Chapter 3: Life and Mind

1.Introduction

1.1.What Is Required To Recognise The Deeper Extent Of Mental Autonomy

My task in the remainder of this thesis is to argue that the autonomous nature of mentalistic explanation presents a stronger constraint on what counts as a satisfactory statement of the relation between the mental and the physical than can be acknowledged within the metaphysical framework of non-reductive physicalism. This means not only that the statement of this relation has to be consistent with the fact that mental concepts are irreducible to physical concepts, but also with the fact that the work carried out by mentalistic explanations is completely separate from the work carried out by physicalistic explanations. To recognise the latter is to recognise the deeper extent of the autonomous nature of mentalistic explanations: mentalistic explanations can carry out the work required of them without implicating the explanatory resources of the physical sciences in their support. In order to recognise the deeper extent of the autonomous nature of mentalistic explanations, therefore, it is not sufficient to find a means of securing the irreducibility of mental concepts to physical concepts as the non-reductivist implies. I suggest that it is *also* required that we find a means of securing the complete separation of the work carried out by mentalistic explanations from the work carried out by physicalistic explanations.

Here is how this will be achieved. I will begin with a detailed defence of the claim that there is an intrinsic connection between the nature of human relationships and the nature of the mental, and I will suggest that an investigation into this connection, as it manifests itself within our everyday explanatory practices, ought to put me in position to successfully accommodate the first requirement. I will suggest, in other words, that the autonomous nature of mentalistic explanation is to be understood in terms of the autonomous nature of human relationships. In making this claim, it will be argued that individuals acquire their repertoire of conceptual skills and mental capacities through learning how to be involved in various types of relationships with other people. Part of this learning process is the process of acquiring the wide range of mental concepts that individuals put to use in their day to day lives, in telling each other what they think about this and how they feel about that, in listening

to each other's stories and news, in discussing new projects with work colleagues, in comforting friends in their troubled times and in expressing sympathy and concern for them, and so on and so forth. The point here is that mental concepts have countless different uses across the different relationships in which individuals are involved, and that using mental concepts is part of the way in which individuals express and direct their interests and concerns in relating to each other. It seems to me that whether mentalistic explanations successfully carry out the work required of them is determined by whether they satisfy the understanding sought by the individuals concerned; once that has been ascertained, there is no further question for these individuals of whether their mentalistic explanations stand in need of support from the explanatory resources of the physical sciences.

But this is only part of my task. What is also required, in addition to this, is a direct attack on the claim that mentalistic and physicalistic explanations converge on the common subject matter of internal behaviour-causing states and events. This will enable me to meet the second requirement. The central strategy in this part of my argument will be to develop a non-causal account of the work carried out by explanations employed in rationalising human action, which will build on the intrinsic connection between the nature of the mental and the nature of human relationships, and which will contrast directly with the causal approach to such explanations integral to the metaphysical framework of non-reductive physicalism. The point of contrast between these different approaches will be crucial for the defence of my thesis: the non-reductivist's account of rationalising explanations maintains, whereas I want to deny, that the notion of causality *must* be built into the concept of what it means for an individual to act for a reason, where the relevant causal relations are implemented at the level of physical processes. It seems to me that if we can maintain that what an individual thinks can completely explain his behaviour, without presupposing that his thinking has to be physically constituted, we will then be in a position to appreciate the deeper extent of the autonomous nature of mentalistic explanations; and hence, we will finally be able to make some suggestions and recommendations as to how the relation between the mental and the physical might be articulated.

One final point before moving on. In the introduction I stated that my intention was to develop the methodological starting point recommended by Baker. Her recommendation is to take our everyday explanatory practices as self-standing and successful in their own right. She claims that mentalistic explanations can meet our explanatory requirements without our

having to presuppose that the mental is related to the physical by way of identity or constitution. Whilst this claim will certainly figure centrally in my overall argument, I will be developing the recommended starting point in a somewhat different direction in order to make a stronger claim. Let me be quite specific about the development I have in mind. To begin with, here is Baker's claim, with which I am in broad agreement:

> The legitimacy of psychology no more depends on particular physical realizations of explanatory states like belief than does the legitimacy of economics depend on particular physical realizations of explanatory states like the national debt. It should be obvious that it does not follow from any of this that beliefs (or the national debt) are realized in, or constituted by, some nonphysical stuff. The point that I am urging is that physical realization or constitution is simply irrelevant to the justificatory, explanatory, and predictive uses to which we put beliefs (or the national debt). (1995: 148).

I will be arguing that not only does the success of mentalistic explanation fail to commit us to accepting that the mental must be related to the physical by way of identity or constitution, but that the success of mentalistic explanation depends on the fact that the mental and the physical are not so related. My central contribution will be made by basing this stronger claim on the point that the identification of the mental with the physical commits us to accepting an account of how mentalistic explanations work which fails to cohere with the way in which they are actually used within the contexts of our relationships. It will be argued that the physicalist's conception of the mental is implicitly tied to an implausible interpretationist account of human relationships, and that a more satisfactory conception of the mental, because it acknowledges the deeper extent of the autonomous nature of mentalistic explanation, can be arrived at through a more accurate account of the way in which we actually relate to each other on an everyday basis. In making this stronger claim, I will be departing from Baker in another respect: whereas she wants to say that mentalistic explanations are adequate if they qualify as causal explanations which do not depend on the truth of physicalistic explanations, I want to say that mentalistic explanations are adequate if they satisfy the (mostly) non-causal understanding sought by each individual within the contexts of his or her own particular relationship.

1.2. Brief Outline Of The Remainder Of This Chapter

In the remainder of this chapter I want to motivate an interpretation of the autonomy of the mental in terms of the autonomy of human relationships, which should provide me with a starting point for completely separating the work carried out by mentalistic explanations from the work carried out by physicalistic explanations. Central to the connection between the

nature of human relationships and the nature of the mental, I will argue, is Wittgenstein's notion of rule governed practices. I will therefore begin in the next section with an outline of some of the salient points concerning the autonomy of human relationships. I will then move on to draw out what I consider to be the important connections between our involvement in various types of human relationships and our involvement in rule governed practices, and then I will deepen this connection by considering the extent to which individuals are dependent on their involvement in relationships with others for their possession of conceptual skills and mental capacities. Finally, I will consider how this connection manifests itself within the contexts of our everyday explanatory practices, and what this implies about the autonomous nature of mentalistic explanations.

2.The autonomy of human relationships

2.1.Introductory Remarks

The first point that needs to be made is that what it means to have the capacity to be involved in human relationships is not something that can be given a straight-forward and exact definition. It is very unclear whether it would be possible to give a finalised and definite list of criteria that individuals must satisfy if they are to be said to have precisely this capacity. Actually, I think that such a list of criteria would not be possible. The reason for this is that our relationships take so many different forms, and are of so many different types, that it would not make sense to suppose that there were a definite list of criteria that individuals had to satisfy, if they were to be said to have the capacity to be involved in human relationships. The possibility of being able to give such a list would presuppose that there were no indeterminacies with regard to the meaning of 'involvement'. But different types of relationships demand different levels and types of involvement, and there are many different reasons and motivations for an individual to become involved in certain relationships, or for him to remain involved, to the same or to a lesser extent, in the relationships in which he is already involved. To be involved in various types of intimate and loving relationships, for example, demands a deep level of emotional commitment and moral responsibility, which is not required of individuals who are involved in various types of business or professional relationships. Different relationships have different types of demands and requirements, which makes the task of regimenting human relationships into one single form impossible,

and it also makes impossible the task of producing a strict definition of what it means to be involved in human relationships that can be generally applied throughout.

It seems to me that this is indicative of the fact that the network of human relationships is irreducible to anything more basic. It is not realistic to expect to be able to define what it means to be involved in human relationships in terms of factors external to this complex network itself. It is the demands and constraints of our relationships themselves that determine what is required of individuals who can be said to have the capacity to be involved in them. Intimate human relationships, for example, demand not only a great deal of emotional and affective commitment from the individuals involved in them; they also demand that each take on a great deal of moral and practical responsibilities toward the other. Working relationships, on the other hand, might not demand this type of emotional and affective commitment, but they do sometimes make certain moral and practical demands on the individuals involved, as a background to the demands which are more specifically work-oriented. It seems to me that not only are different types of relationships constituted by different types of demands, but that within any one type of relationship there are certain demands which might also be common to many other relationships. It is therefore difficult to isolate any one type of human relationship in order to exactly specify which demands are uniquely constitutive of it. It is more likely that the various types of human relationships in which individuals are involved are interdependent in varying degrees, and that the demands which are partly constitutive of one type of relationship are at the same time demands which are partly constitutive of other types.

The autonomy of human relationships manifests itself most clearly in the fact that the demands which are partly constitutive of our relationships are demands which can only come into view for us through our involvement in the relationships themselves. There are two important points which can be taken from this: first, it is not possible to make sense of the demands of our relationships on us except from within the contexts of our involvement in these relationships themselves; second, the interdependence of the demands of some types of relationships on the demands of others undermines the possibility of reconstructing the demands of our relationships at a more basic level of explanation. The problem is that in order to get these demands into view a great deal of intentionalistic and moralistic language has to be in place, but this does not seem possible except within the contexts of the relationships themselves.

In developing these points, I want to focus on the idea that although the demands presented to an individual in the various relationships in which he is involved are not specifiable in any strict form, this does not detract from the individual's understanding of what is required from him if he is to be able to respond appropriately to those with whom he is involved. I want to stress that the individual's understanding of how he ought to respond in any given situation could not have been the product of some form of training that successfully drilled into him a definite list of demands and a definite list of responses. Rather, what the individual knows that enables him to cope with the situations in which he finds himself, to respond to the demands which are presented to him in his involvement with others, could only have been acquired through experience of living out his life with other people on a day to day basis. In effect, I will be claiming that although the various practices in which individuals participate throughout the normal course of their lives are autonomous, they must also be characterised by a degree of indeterminacy, which stems from the lack of strictness and rigidity in many of our shared routines and activities.

This indeterminacy should not be considered in a negative light, however, as a failing that ought to be completely eradicated, or minimally improved upon. It is necessary to sustaining our involvement in many of our relationships that this indeterminacy is retained as it is. One reason for this is that any attempt to forcibly inject a greater amount of regimentation and rigidity into the routines and rituals of our everyday lives would result in the loss of the depth and diversity in our relationships, which is considered to be important to our own sense of who we are. The variety in our relationships, or what comes to the same thing, in our shared routines and activities, is fundamentally important to our own sense of personal identity. It is precisely because there are such differences in our relationships that certain individuals become more important than others to sustaining our sense of who we are, and to sustaining our sense of personal worth. Dilman puts the general idea quite neatly:

> A person is who he is in those relationships which mean most to him, relationships which engage his deepest loyalties, obligations, love and gratitude. The first relationships of this kind are the network of family relationships into which he is born and in which he finds growth. Later come friendships, his work and colleagues, the demands these make on his moral resources, the interests and loyalties they engage, what he gives to them, his sexual loves, marriage, children, and other commitments. But what he learns in and from his earliest relationships, who he becomes in them and what he comes to be like, largely set the pattern for what he makes of these later relationships. (1990: 208).

My suggestion, therefore, is that since human relationships are autonomous, in the sense that their demands cannot be redescribed at any level other than that of our involvement with others in everyday activities, they must also be characterised by a certain amount of indeterminacy, in the sense that there are no strict definitions of how an individual ought to respond to those with whom he is involved. If autonomy is to be understood as a feature of the complex web of human relationships, and hence of the various activities in which we engage, then it must be balanced with the degree of indeterminacy that is also a feature of many of our everyday activities. I suggest that we ought to think of this balance in terms of the balance that Wittgenstein sets out to achieve, between the autonomy and indeterminacy that characterise the rule governed practices in which individuals participate in living out their everyday routines. My deeper reason for appealing to Wittgenstein's rule following considerations, other than the fact that they suggest a way of balancing this autonomy and indeterminacy, is that I believe there to be an intrinsic connection between what it means to have the capacity to be involved in human relationships, and what it means to have the capacity to be engaged in rule governed practices: to have the capacity to be involved in human relationships requires having the wide array of conceptual skills and abilities that individuals develop through learning how to participate in rule governed practices.

2.2a.An Interpretation Of Wittgenstein On Rules And Relationships

Wittgenstein's rule following considerations are typically thought to be an attack on the platonistic conception of rules which, when worked out to its logical conclusion, turns out to be self-refuting. No doubt there is some truth in this. But it seems to me that the target is rather any conception of rules, not just of the platonistic variety, which holds that the identity of the rule is logically prior to the identity of the various routines and activities constitutive of our everyday practices. What motivates this conception of rules is that it seems to offer the only means of retaining the assumption that the rule itself determines its own applications, autonomously and independently of those applications which strike the individual who is following the rule as correct. According to the target conception, this assumption finds expression in the idea that the rule ought to be conceived as something like a universal formula, whose applications have already been settled in advance of any application which might be made of it. This indeed seems to guarantee autonomy, since the rule is already supposed to have determined its own applications. But the difficulty with this is that it presupposes that the identity of the rule can be fixed from the outset without taking into

consideration the indeterminacy inherent in many of our practices: in certain situations, where it is not possible to discern the required degree of rigidity in our activities, the mechanical application of such a pre-set universal formula is simply not going to work. The indeterminate nature of our practices presents the individual with a conflict between the absolute rigidity of the rule and the unformalisable activities in which he is engaged, and it looks as if he will in fact have nothing to appeal to, in deciding how to go on in any particular case, other than his own possibly conflicting interpretations of the rule. This is the difficulty which Wittgenstein dramatises, where he has his interlocutor ask: "But how can a rule show me what I have to do at *this* point? Whatever I do is, on some interpretation, in accord with the rule" (1967: § 198).

Wittgentsein's point seems to be that the autonomy of rules cannot be retained if it is conceived in the manner of the target conception, since that conception can be seen to give rise to the possibility that the rule fails to determine what is required if it is to be applied correctly. Confronted with a situation in which the mechanical application of the logically pre-set rule seems to be out of place, the individual is then left to his own devices in deciding how the rule ought to be applied in this case. The problem is that the activities surrounding many of our rules cannot be formalised in the manner required to support the idea that the identity of the rule is logically prior to the identity of our practices. Indeed, this conflict is heightened in cases where the activities surrounding our rules leave certain of their applications open to question. Suppose, for instance, that there is a rule stating that items which might be used as offensive weapons cannot be sold in hardware stores to anyone under a certain age. It is perfectly conceivable that, regardless of how precisely this rule is stated, situations will arise in which it is not very clear whether the trader has correctly applied this rule or not. Whilst it is clear enough that selling a knife to a youth would be an instance in which the trader has gone against the rule, it is not so clear that selling a strip of wood, which might be sharpened into a spear or used as a baton, would be such an instance. In this case there is no clear answer to whether the trader would be going against the offensive weapon rule by selling the strip of wood to the youth. And as Wittgenstein is happy to acknowledge, "rules leave loop-holes open, and the practice has to speak for itself." (1979: § 139).

Thus, Wittgenstein's problem is with the approach to autonomy which forces rules to be conceived in such a manner that their applications require complete determinacy throughout our practices, when this simply cannot be granted, given the lack of rigidity that surrounds

many of our everyday activities. The problem can be brought into focus by considering the fact that there are certain types of rules whose applications cannot be settled in advance as the target conception assumes, since there are always going to be situations in which the mechanical application of the rule is upset by the lack of rigidity in these activities. The idea seems to be that unless the identity of the rule is fixed within the practice itself, thereby leaving room to accommodate the indeterminacy inherent in the activities surrounding the rule, it looks as if it will not be possible to retain the general assumption that the rule determines its own applications autonomously and independently of the applications which strike the individual as correct.

I think Wittgenstein's suggestion is that this balance can be retained if it is held in place by the *attitudes* that individuals take toward the various activities in which they are disposed to engage. The idea is that the relevant attitudes are cultivated in the course of bringing the individual to *see* that certain types of action *must* be performed if he is to follow certain rules correctly; the resulting generality in these attitudes is responsible for carrying the individual confidently into situations which had not been explicitly mentioned in the statement of the rule itself. The balance between autonomy and indeterminacy is achieved through the realisation that it is only through living out the routines and activities constitutive of our practices that the individual develops the skills and abilities to cope with the demands of the situations in which he finds himself. It is these very same skills and abilities that enable the individual to understand what is required of him if he is to continue to follow certain rules *correctly* in each new situation, even if the new situation seems to present the individual with demands within a set of circumstances which he had previously not encountered. The autonomy of rules is preserved in that the generality in the individual's attitudes, which are expressed through exercising the relevant conceptual skills, commits the individual to accepting only a certain type of response as correct.

The type of response judged to be correct may or may not already form part of the everyday routines and activities in which the individual is engaged. The individual draws on his repertoire of conceptual skills and abilities already built up through his day to day experience, on his basic practical know-how, to determine the correct response in this new situation. But given the lack of rigidity in many of our activities, there need be no strict definition of the correct response in this new situation. If the correct response in this new situation is not one with which the individual is already familiar from a different context, it could be that, when

confronted with such an atypical case, the individual might have to come to his own informed decision based on what he has already learned through previous experience; in other situations, he might have to rely on the explicit judgement of others, who may have had the benefit of a greater and more varied experience. But the important point is that once the identity of the rule is fixed within the practice itself, this amount of indeterminacy can be tolerated, in so far as it is indicative of the lack of rigidity in certain aspects of the activities in which individuals are engaged. It would be a mistake, however, to think that toleration of a degree of indeterminacy in certain situations could licence toleration of complete indeterminacy throughout our practices, as would have to be the case to warrant an appeal to the agreement in judgements of others in the individual's linguistic community as the only means of settling the correctness of the individual's claim to have mastered any rules whatsoever.

It seems to me that part of what explains the indeterminacy in certain types of rules is the fact that an individual's engagement in the relevant practices is not sharply separable from his involvement in various loosely structured activities and routines with other people. Our rule governed practices are shaped and structured through the different ways in which individuals relate to each other, which means that certain types of rules and demands are going to be less rigid and exact than others. For neither the complex web of human relationships, nor the array of rule governed practices in which we engage, can be treated independently, as separate going-concerns.

The autonomy and indeterminacy in our practices cannot be grasped independently of grasping the autonomy and indeterminacy in our human relationships, and the balance between the autonomy and indeterminacy is held in place in both cases by the attitudes that individuals learn to take toward the routines and actvities in which they engage *with each other*. We refuse to accept certain applications of rules as correct, just as we refuse to accept certain reponses to others as appropriate. But in many cases there is no strictness in our definition of what counts as a correct or appropriate response, and there is no formalisation of these responses which could be given in advance to cover each and every case. What this suggests is that the indeterminacy in certain of our rules, or in the demands of some of our relationships, is to be understood in terms of the lack of rigidity in these activities and routines, whence it might happen that a specific doubt as to how a particular rule ought to be applied in a certain situation, or as to how a particular person ought to be treated on a specific

occasion, has to be settled by making an informed decision, or by appealing to how other more experienced people would be inclined to respond in that same situation.

2.2b. Some Comments On Kripke, Baker And Hacker, Malcolm and Williams

Although the appeal to others certainly has its place in settling the application of certain rules in specific contexts, it seems to me that Kripke's (1982) appeal to how others would be inclined to respond, as the *only* means of settling the correctness of the application of every rule, betokens a failure to appreciate the need to retain the rigidity in certain of our activities, such as counting, measuring, telling the time, building bridges, drawing maps, and so on, without which our lives could not continue to function in the regular manner that they do. Whereas it is perfectly acceptable to tolerate a degree of indeterminacy in some of our rules, where the identity of those rules is fixed within the context of activities which demand less exactness and strictness from the individuals involved, it would be a mistake to think that this indeterminacy could characterise every aspect of our practices, putting individuals into the position where every application of every rule had to be checked for correctness against the inclinations of others in their linguistic community. But this is precisely how Kripke interprets Wittgenstein on this matter. Here is what he has to say:

> Any individual who claims to have mastered the concept of addition will be judged by the community to have done so if his particular responses agree with those of the community in enough cases (1982: 91-92).

Kripke's community view of rule following is developed in response to the sceptical point that there is no fact of the matter which determines whether an individual is currently applying a rule in accord with his previous linguistic intentions; it is proposed as a direct return to the sceptic's challenge that there is no determinate answer to the question of whether an individual is now justified in applying a rule in one particular way rather than another. For if we suppose that the current application of the rule takes place within an entirely new and hitherto unencountered situation, and if we ask whether the individual is now applying the rule as he had previously intended it to be applied in this situation, then we seem compelled to agree that the question has no definite answer. The problem is that since the individual applied the rule a finite number of times on previous occasions without giving himself explicit instructions as to how the rule should be applied in this new situation, it seems to follow that any application of the rule in this situation could be made out to be compatible with his previous applications of the rule. This is Kripke's famous example:

> Let me suppose...that '68+57' is a computation that I have never performed before...I perform the computation, obtaining, of course, the answer '125'...Now suppose I encounter a bizarre sceptic...Perhaps, he suggests, as I used the term 'plus' in the past, the answer I intended for '68+57' should have been '5'!...After all, he says, if I am so confident that, as I used the symbol '+', my intention was that '68+57' should turn out to denote 125, this cannot be because I explicitly gave myself instructions that 125 is the result of performing the addition in this particular instance. (1982: 8).

Kripke deals with this problem on Wittgenstein's behalf by appealing to the agreement in the judgements of the linguistic community to provide the criterion for what counts as a correct application of the rule. Faced with the possibility of radical indeterminacy, which threatens once the sceptic's point is accepted, it seems that the individual must resort to how others would be inclined to apply the rule to determine whether his own application of the rule is correct. But as Baker and Hacker (1992: 171-2) complain, the problem with the community interpretation is that since it is committed to the idea that the criterion of correctness is not provided by the rule itself, but by an external agency, it is guilty of abrogating the *internal relation* between a rule and acts in accord with it. Their central point, with which I am in agreement, is that the individual's training into our linguistic practices instils in him the capacity to discern a correct application of the rule from an incorrect one, and what he thus discerns is not that his judgements agree with the judgements of others, but that *the rule itself demands that a particular action must be performed if it is to be applied correctly*, and that in applying the rule he is thus acting in accord with the already existing practice of following the rule.

Baker and Hacker use this point to argue that there is no need to think of rule governed practices as requiring more than one individual. Which amounts to the claim that if other people are not required to settle the correctness of an application of a rule, then they are not required at all. On the strength of the fact that there is an internal relation between a rule and acts in accord with it, Baker and Hacker argue that Wittgenstein's reference to a practice is not *necessarily* reference to a practice which involves a multiplicity of agents. And given this, they argue that there is no logical incoherence in supposing that a Robinson Crusoe could establish a novel rule governed practice and then engage in it successfully. But whilst I think they are correct to insist that an isolated individual could establish his own practice and engage in it without losing sight of what counts as correct applications of his rule, I am not convinced by their much stronger suggestion, that it is possible for a radically isolated

individual, one who has been isolated throughout his entire life, one who has had no contact whatsoever with others, to create and sustain his own rule governed practices.[1]

Against the idea that there can be such solitary rule followers, Malcolm (1995) argues that there is an important sense in which other people are in fact indispensable to our rule governed practices. Malcolm defends this claim in the course of criticising Baker and Hacker for not being able to answer what he refers to as the *hard question*. The hard question is this: what decides whether a particular step taken, a particular application made, is or is not in accordance with the rule? It is important to note that Malcolm's complaint is not with the idea that the rule and its applications are internally related, but with what Baker and Hacker take to be the implication of this idea, namely, that the rule itself, and nothing but the rule itself, determines which steps are in accord with it. The point is that if the rule and nothing but the rule determines what is correct, then it would follow that when a rule is given, so must its extension be given. But the central problem, by Malcolm's lights, is precisely that when a rule is given, its extension is *not* given. So what is required to answer the hard question, if Baker and Hacker have not been able to do so? Malcolm thinks that since it is a mistake to say that when a rule is given, its extension is given, the hard question can only be answered by recognising the importance of a framework of quiet agreement to provide the rule with its extension:

> In asserting that 'the rule and nothing but the rule determines what is correct', Baker and Hacker do not seem to give sufficient weight to Wittgenstein's insight that a rule does not determine anything except within a setting of quiet agreement. If you imagine that no longer existing, you become aware of the nakedness of the rule. The words that express the rule would be without weight, without life. A signpost would not be a signpost. A rule by itself determines nothing. (1995: 149-50).

According to Malcolm, the nakedness of the rule is evident in the fact that different people, with similar training and equal intelligence, could form different extensions in accordance with the same general expression of the rule. So, without a framework of quiet agreement, our understanding of what rules are would disappear. Consider one of his examples. Suppose that you arrive at the junction of a busy London intersection. If there were no agreement among drivers as to which direction to turn in following a sign, signs would no longer

---

[1] Baker and Hacker claim that Wittgenstein conceived the possibility that an individual could be "acquainted only with language games he played with himself" (1992: 175). This implies that an individual could invent his own language games despite not having been involved in shared language games at any other time. I am doubtful that this can be the case. It seems to me that an individual could certainly invent language games, which only he played, but I am not sure that he could be acquainted only with language games he played with himself, for this implies that he need never have been acquainted with shared language games before he could invent his own.

function as signs and chaos would ensue. Malcolm is careful to point out that drivers do not consult each other's opinions to decide whether the sign indicates *this* direction or *that*, but he does hold that unless there is a framework of consensus of action and reaction, there would be no such thing as rules, signposts, and so on. The agreement which is required is not an agreement in the opinions of all the drivers as to which way to go, but it is an agreement in the way we are all trained to react to signposts. Malcolm thinks that Baker and Hacker's insistence that "it is the rule and nothing but the rule which determines what is in accord with it" blinds them to the possibility of widespread disagreement in the application of rules, and as such leads them to underestimate the significance of agreement in action for the concept of a rule.

It seems to me that Malcolm makes an important point against Baker and Hacker, but he puts it rather misleadingly in saying that "a rule by itself determines nothing". For, when Wittgenstein claims that a rule is not an extension, he might only be objecting to the platonistic conception of rules outlined earlier, and not to the idea that there is a perfectly ordinary, non-platonistic, sense in which the normative aspects of rules are autonomous. Malcolm seems to have recoiled too far from the platonistic conception of rules, leading him to say that when a rule is given, its extension is not given. He appears to be thinking of a rule as being in itself normatively inert, unable to determine what is to count as a correct application of it until it is dressed up within a framework of quiet agreement. Malcolm could have put his point by saying, *not* that without general agreement the rule determines nothing, but that without general agreement there would no rules at all. He does seem to suggest this himself, however, when he says that without agreement a signpost would not be a signpost, but this specific way of putting the point does not seem to have the same consequences. It seems to me that if Malcolm had settled with this formulation, he would have been able to avoid the problem that gave rise to the hard question. For if we put the point this way, we do not have to assume that there can be such a thing as a rule whose applications are completely undetermined, until it has been brought into a framework of agreement in action. The most we have to assume is that without such agreement, there would be no rules at all. This way of putting the point is perfectly consistent with saying that an individual might create his own rules that only he follows, as in the case of Defoe's Crusoe, but it is inconsistent with saying that this can happen in the case of an individual who has been isolated throughout his life, which is what Malcolm seems to have been getting at anyway.

But regardless of how unusual the case seems, it might be objected that there is no logical contradiction in the supposition that the radically isolated individual could possess conceptual skills. It might be argued that the radically isolated individual could have acquired the latter through some magical or super-natural process, perhaps through swallowing pills whose effect is to make him a competent rule-follower, or through being struck by mysterious language-inducing bolts of lightning. Given these possibilities, the objection might continue, we have no reason *a priori* to deny that the radically isolated individual could be in possession of a language, and this presents an apparent problem for my claim that an individual acquires his repertoire of conceptual skills and mental capacities through learning how to be involved in relationships with other people.

However, as Meredith Williams (1991: 119) points out, the force of this objection is weakened by the fact that an individual's conceptual skills could not have been acquired spontaneously in these ways, because to attribute such skills to an individual at any point in time presupposes their prior duration through time. Or in other words, since the individual's conceptual competence in following sign-posts or continuing arithmetical series is manifest over time, it follows that in attributing such skills to this individual we are committed to assuming that he has the know-how to act in an appropriate manner in the future, and that he has either already acted in an appropriate manner on previous occasions, or that he has a history which would lend some warrant to our judgement that he is at last beginning to grasp the rules. To specify a particular point of time at which the individual suddenly acquired his conceptual skills, and at which he suddenly displayed behaviour which we would call linguistic, without filling in details about the history of the individual's current conceptual performance (which would arguably involve details about his initiation into a shared way of living), is therefore unintelligible.

I think that this is certainly a damaging point against the idea that the radically isolated individual could possess conceptual skills. But it might be replied that this point only threatens the latter idea on the assumption that the magic pills and mysterious bolts of lightning bring about the acquisition of language spontaneously. For if we were to assume, on the other hand, that the effect materialised only after a longer period of time, during which the radically isolated individual began to act in a manner apposite to acquiring the relevant history, then the point would lose much of its force. All that would be required would be that the individual got into the habit of scratching - -- --- ---- in the sand, for instance, and that he

eventually began to use such patterns, along with variations on them, to represent and keep a tally of the fish he had speared that day. But once we have conceived this possibility, we seem to be only a short step from dispensing with magic, and granting that the radically isolated individual could have brought about this state of affairs by himself. On this assumption, the case against the radically isolated language user becomes unclear, and it becomes equally unclear whether the issue can be definitively settled one way or the other.

So instead, it will be worthwhile considering whether other people are in fact indispensable to our having the conceptual skills and mental capacities *we* do, which is the point I really need for my overall argument anyway. The point is that language is a lived phenomenon, and the indispensability of others to our linguistic practices is connected with the fact that our language is bound up with our emotional and affective natures, with the basic needs, interests and concerns we express in a variety of different ways in the course of our lives together. Certainly, this does not logically exclude the radically isolated individual from possessing *a* language; but it means that *if* he can be said to posses a language at all, the type of history we would be obliged to attribute to him would be such that his language would fail to have many of the features characteristic of *our* language. The upshot of this is that the type of practice that would be accessible to the radically isolated individual would not be the type of practice that is accessible to individuals who are involved in a shared way of living such as ours. The type of situation in which the radically isolated individual could find himself would be significantly different from the type of situations in which individuals in relation could find themselves, given that the situation would be entirely of the individual's own making. It would not be shaped in any sense by the needs and demands that shape the situations in which individuals in relation find themselves where, for instance, there is some concern for the interests and the welfare of others, where there is a degree of respect for persons and their perspective, where there are fears that some people will harm them, and so on, since it would not be possible for any other individual to exert the influence of their presence in shaping his conception of the situations in which he finds himself.

This manifests itself most clearly in the fact that the radically isolated individual could not have the same mental concepts as we do, if indeed he can have any, since the uses of our mental concepts are shaped by the various ways in which we relate to each other. The type of practice accessible to this individual would thus lack many of the features necessary to support the claim that he had the same array of emotional and mental capacities as the

individual who had the benefit of being involved in relationships with others. The radically isolated individual's conception of pain and fear, for instance, would not be the same as our conception of pain and fear. The simple reason for this is that the different ways in which other people are involved in our lives contributes to our understanding of what it means to be in pain or to be afraid, which means that the radically isolated individual would necessarily lack our understanding of these concepts. An individual in pain is an individual towards whom sympathy is an appropriate attitude to have, and an individual who is afraid is an individual towards whom comfort is an appropriate attitude to have. The radically isolated individual's conception of pain and fear would necessarily lack these normative implications, since their meanings would be tied exclusively to the contexts of his own individualistic world. So the radically isolated individual would not be in command of important aspects of *our* mental concepts since he would never be in the position to respond to the type of demands that are presented to individuals through their involvement in human relationships. Nor, for that matter, would he ever be in the position to use mental concepts as part of the ways in which individuals relate to each other, to make promises or to confess sins, to express gratitude or to declare love, to frighten, humiliate or to humble.[2]

2.3.The Inner And The Others

By appealing to certain features of Wittgenstein's rule following considerations, I have been trying to suggest that individuals can only be said to have the conceptual skills and mental capacities they have through their involvement in a particular way of living their life with each other. On the strength of this, I claim that since the process of learning to relate to others must involve acquiring the conceptual skills which alert individuals to the demands of these relationships, it is this very same process that is responsible for giving content and structure to each individual's own personal psychology. In this section I will try to show that this claim is not as implausible as it might appear, that it does not threaten the deeply complex and often intensely private nature of the inner. I will do this indirectly: first, by touching on an aspect of Wittgenstein's (1992: 62-3) idea that the inner and the outer are related logically, and that the relation between the inner and the outer is itself part of the concepts; and second, by dealing with an objection to this way of approaching the inner in general.

---

[2] These ideas will be further developed in the next chapter.

Part of what Wittgenstein is trying to get at here is that that there is a logical relation between the inner and the outer which is fixed by the concepts which are constitutive, not only of the inner experiences that the individual can be said to have, but also of the outer circumstances of which certain of his inner experiences can be said to be experiences.[3] An example might help to illustrate this point: an individual's experience of being struck by the thought that his door has been left unlocked is an experience which is intrinsically concept involving, such that the individual could not be said to have this type of inner experience if he could not be said to have the conceptual skills which are drawn on in having it. The concepts which are constitutive of his inner experience are the very concepts which are constitutive of the fact that his door has been left unlocked, and as such the concepts involved serve to fix a logical relation between the inner experience and the outer circumstance of which the inner experience is an experience. The important point for my purposes is the former: since the inner experience is intrinsically concept involving, it must be a logical precondition of having this experience that the individual has the conceptual skills which are drawn on in having it. In other words, or so I claim, it must be a logical precondition of having this experience that the individual has the capacity to be involved in the various types of relationships, in terms of which the fact that his door is unlocked strikes him as particularly salient.

But on the surface of things, this idea might look fairly implausible. Or at least, it might look as if it is not able to provide convincing support for the claim that an individual can be said to have his own personal psychology only through his involvement with others. For the obvious objection to make against this idea is that it forces such a strong dependence relation between the inner and the outer, that it not only fails to do justice to the density and complexity of inner experience itself, but that it is also in danger of refusing to fully acknowledge that the individual has his own personal psychology. Murdoch (1992), for instance, argues in this vein, that any attempt to explain the nature of the inner in terms of logically structured practices leads to the denial of inner experience, and ultimately to the loss of the individual. Her concern seems to be that if the logical precondition of having inner experiences is that the individual is involved in a variety of shared practices, then the individual must lose his own identity to take on the larger identity of the people with whom he is involved. Murdoch's

---

[3] For further discussion of the internal connections between the inner and the outer, see McDowell (1991: 159-160), and for a discussion of how these ideas tie up with the private language argument, see McDowell (1989). The common thread between these discussions is the idea that it makes no sense to think of inner experience as resulting from the application of concepts to previously unconceptualised occurrences; rather are inner experiences already conceptually constituted as the inner experiences they are.

problem is that inner experience is simply too complex, too intensely personal, and too fluid, to be analysed away in terms of the individual's involvement in such practices, which are devoid of the personal values, moral and aesthetic concerns, and so on, that must be in place if the individual is to be credited with his own personal psychology. Murdoch writes:

> Experience has layers. Here the intense lively privacy of the individual's 'inner life' presents itself as something not to be analysed away...[subjecting] an unconquered field to a particular neater clearer account...Must a particular technical mastery be a 'logical' condition of someone's having a particular experience? (What indeed is logic doing here!)." (1992: 278).

This is an important point, for it is most likely that the individual's experience of being struck by the thought that his door is unlocked will not be as simple as I seem to have suggested. It is most likely that this experience will also be tainted with various fears and worries, that it will cause the individual to panic about his house being burgled, that it will make him critical of his negligence, and so on. As Murdoch points out, this type of inner experience will involve innumerable considerations, in the sense that it will leave too much out to say that it is simply about 'the door's being left unlocked', and it will also involve certain value judgements, in the sense that the individual having this experience will take up some sort of attitude to the fact that his door has been left unlocked. But it seems to me that this can be explained by saying that it is precisely because the individual is involved in the types of relationships which alert him to this fact in the first place that his inner experience takes on the complex character it does. It is precisely because the individual lives in a society in which the activities of criminals and thieves force us to be security conscious that his unlocked door strikes him as particularly salient, and it is this very same set of considerations that explains why the individual's sudden realisation that his door is unlocked will arouse all sorts of worries and fears about his house being burgled. I think this provides us with a way of dealing with Murdoch's point that inner experience involves moral judgements and personal concerns. But whereas Murdoch thinks this means that logic ought to be expelled from the inner, I think it means that the identity of the inner experiences that an individual has cannot be sharply separated from the identity of the relationships in which he happens to be involved.

Before moving on, I want to suggest that the type of worry raised by Murdoch is better aimed at positions where emphasis on social structure gives rise to a rather negative construal of the inner. Hampshire (1976) proposes an account of the dependence of inner feeling on the individual's awareness of social constraints, which might help to illustrate the worry.

Hampshire's idea is that an individual's ability to have inner feelings is logically dependent on his ability to exercise restraint on his natural dispositions toward certain patterns of behaviour, and his ability to exercise restraint on these dispositions is logically dependent on his ability to identify the patterns of behaviour as being appropriate or inappropriate within certain situations. Through becoming aware of the contrast between how he is naturally inclined to respond in certain situations, and how he is socially required to respond in these situations, the individual learns to exercise restraint on his behaviour. And the felt inclination, which remains when the individual restrains himself in these situations, is thus identifiable only in terms of the patterns of behaviour toward which it is a restrained inclination. Hampshire expresses this dependence in the following way:

> in the particular case of feeling, the inner life of the mind is to be understood as a development of something more primitive in every man's behaviour, of which it is the residue and the shadow (1976: 73).

The point might be explained by saying that if an inner feeling is that which remains as a 'residue' or 'shadow', through the exercise of restraint, then the inner feeling must be constituted by the very concepts which pick out the patterns of behaviour that have been restrained. So it follows from this that the identity of the inner feeling is not logically separable from the identity of the patterns of behaviour toward which it is a restrained inclination. At first glance this might seem plausible enough, especially if we restrict our attention to anger and fear, and if we think of the latter in terms of lashing out or running away: the inner feelings which remain when restraint is exercised *are* logically secondary to such natural patterns of behaviour. But now Murdoch's concerns must come to the fore, since this approach does seem to throw a negative, and somewhat restrictive, mantle over the inner in general. To begin with, it is too narrow as it stands. There are specific types of inner feeling, like the feeling of being safe, or the feeling of having been somewhere before, which are not restrained inclinations to act, but which rather emerge in their own right as the individual learns to use language spontaneously in new and interesting ways. There are no natural behavioural dispositions which are logically prior to these types of inner feeling, and which serve as their developmental basis in the way that hitting or cowering do in the case of anger or fear.[4] And furthermore, as Dilman points out:

---

[4] Another such example- the feeling that everything is unreal. Wittgenstein (1980a: §§ 125-6) gives the following account: "The feeling of the unreality of one's surroundings...Everything seems somehow not *real*...But why do I choose precisely the word "unreality" to express it?...I choose it because of its meaning...The fact is simply that I use a word, the bearer of another technique, as the expression of a new feeling. I use it in a new way."

there are other emotions, such as grief, guilt, shame and remorse, where the natural inclination is to hide, to turn into oneself, to seek solitude. To check these inclinations in oneself, if one wishes to hide one's feelings, one has to affect indifference, brazen it out, put on attitudes, pretend. Here what one goes on feeling is not a 'residue' or 'shadow'. For though there is a guilt of which defiance is a common secondary reaction, defensive and self-protective in character, the feelings I have mentioned mostly thrive in an inner life- 'inner' in contrast with a life of action. They have their home in such a life; on the whole they incline the person to reflect, reminisce, criticize himself, rather than act. (1987: 53).

To this it can be added that even the apparently paradigmatic cases of anger and fear do not neatly fit into the mould of 'residue' or 'shadow'. Rather than being inclined toward natural patterns of retaliatory behaviour, it is quite likely that an angry individual who has had good moral guidance in his early life will be inclined to seek a means of defusing the situation, or distancing himself from the the source of his anger if possible. Also, a frightened individual may be naturally inclined to talk to someone, or in extreme cases to withdraw into himself rather than act. Here we can begin to appreciate the fluidity and the intense lively privacy of the individual's inner life, which Murdoch is quite correct to emphasise, and which is in danger of being lost in Hampshire's approach to the inner.

That said, I think that Hampshire is certainly correct to stress the importance of social constraints imposed on individuals as that which gives shape and structure to the inner life. So perhaps his good point could be put simply by saying that since the individual learns to exercise restraint, and indeed to express himself in new and interesting ways as he becomes involved in various types of relationships with others, it is in terms of the demands of these relationships that the individual's inner life comes to take on the shape and structure it has. However, that there are such logical preconditions of having inner feelings and experiences does not have to be understood as being in any way restricting or crippling. It is certainly a logical precondition of having the array of inner feelings just mentioned that the individual has mastered the relevant conceptual skills, but this is precisely what enables him to go on to have intensely private, and often deeply moving, experiences which simply cannot be analysed away in terms of a 'particular neater clearer account'. So despite Murdoch's worries, which would certainly hold with respect to certain 'negative' accounts of the inner, it is not implausible to say that the individual's own personal space is inextricably linked with his capacity to be involved in various types of relationships with others.

2.4.Indeterminacy

The connection between the nature of the mental and the nature of human relationships manifests itself most clearly in the indeterminacy that characterises our everyday explanatory practices. Part of what it means to say that the mental is indeterminate is that mental concepts are used within the contexts of our relationships in expressing the interests and concerns that we have toward each other. To the extent that our involvement in our relationships is not rigidly structured, the use of mental concepts in these relationships must capable of adapting to cope with this lack of rigidity. So it seems that the use of mental concepts cannot be fully determinate, in the sense that their use in every possible case can be settled in advance, without losing the flexibility and adaptiveness constitutive of their use within the contexts of human relationships. The important point here is that the indeterminacy in the use of mental concepts is indicative of the fact that there are significant differences in the ways in which we relate to each other. There is a variety in our relationships, and the continued existence of this variety, which would be lost if mental concepts were to lose their flexibility, is important in our lives.

If an individual related to his loved ones as he related to the ticket inspector on the train, for instance, there would be no deep relationships in his life which he considered to be special in themselves, and worth working hard at when things became difficult or strained. There would be little difference between relating his terrible news to his family or relating it to the postman who greets him each morning with his mail. It is important to individuals that different people are involved in their lives in different ways; it is the fact that an individual's loved ones are who they are in his life that he seeks their company, that he wants them to be there, at such significant times. So I think that what it means to say that our mental concepts are indeterminate is that their use is shaped in such a way as to cope with the differences in the depths and qualities of our various relationships. If it is part of the use of mental concepts to express our interests and concerns in our relationships, then the ways in which they perform these tasks will display a lack of uniformity across the variety of ways in which individuals relate to each other.

The glance of a loved one, for instance, might be loaded with feeling, and the affective responses it evokes in the individual it is directed at will be intrinsically concept involving. It is through these concepts that the glance is understood to have the significance it has; but that

it has this significance, and that it evokes the intended response, is inseparable from the depth and quality of this particular relationship. The significance of the glance might, however, be unclear to a stranger to this relationship, whose non-involvement in it may leave him unmoved. There might be no straightforward way in which to bring the stranger to see the significance of this glance, other than by filling in some of the background and the context of the particular relationship in which it occurred. But exactly which details would have to be given to remove the stranger's uncertainty over its significance cannot be determinately settled in advance for every possible case, without assuming an artificial simplification of our intensely complex relationships.[5] It rather depends on the individual one is trying to convince, on how well he relates to others, on his social and interpersonal skills, on how quick he is on the uptake, and so on. Wittgenstein captures this point in his remarks that:

> It is important, for instance, that one must 'know' someone in order to be able to judge what meaning is to be attributed to his expressions of feeling, and yet that one cannot describe what it is that one knows about him. (1992: 89-90).

> What one acquires is not a technique; one learns correct judgements. There are also rules, but they do not form a system, and only experienced people can apply them right. (1967: 227).

In learning to relate to others the individual does not receive a well defined instruction manual, which he can either memorise in full, or consult from time to time, in order to bring the strict rules to bear on particular cases; rather are the individual's skills and abilities, his capacities to be moved and touched by the feelings and experiences of others, gradually built up throughout the course of his life. The individual can be said to acquire a knowledge of human nature through his experience of living his life with others, through the help and guidance he receives from those who may be more sensitive than he is, through his own trial and error, and so on. But what the individual comes to know is not susceptible of codification in a form that would allow the transmission of this knowledge to an individual who altogether lacked the experience of living out his life in the company of others. For that to be possible, we would have to envisage a higher degree of regimentation throughout our relationships than we can readily grant without losing sight of the individual's own private and personal space. But individuals could never be in the position to be credited with their own personal psychology, if it were not possible to see in them the different traits and qualities which

---

[5] As Marie McGinn (1998: 55) points out, the indeterminacy in our psychological concepts is part of the way in which they are bound up with our complicated lives together, and "the uncertainty that enters into our everyday human relationships...[is] an indication of how intricate our life together has become."

gradually emerge out of their involvement in their own complicated webs of human relationships.

3.Conclusion

The autonomous nature of mentalistic explanation comes to this. Mental concepts are variously used in expressing the specific interests and concerns that individuals have within the contexts of their relationships with each other. Since these interests and concerns vary according to the type of relationship in question, what counts as an adequate or successful mentalistic explanation is assessed relative to its ability to satisfy the understanding sought by each individual within his or her own particular relationship. An interest in the underlying causal processes responsible for bringing about another person's bodily movements on a certain occasion is external to the interests individuals have within the context of their everyday relationships, and as such mentalistic explanations can successfully meet the demands *we* make of them without having to implicate the explanatory resources of the physical sciences. What has to be appreciated here is that the nature and the structure of our explanatory practices are inseparable from the way we live our lives together, and that since the standards by which we assess the adequacy and success of mentalistic explanations are therefore generated from within the contexts of our everyday relationships with each other, mentalistic explanations do not stand in need of support from the more specialised explanatory practices of the physical sciences.

To sum up: my claim is that in order to recognise the deeper extent of the autonomous nature of mentalistic explanation it is necessary to completely separate the work carried out by mentalistic explanation from the work carried out by physicalistic explanation. The aim of this chapter has been to provide a framework for effecting this separation by arguing that mentalistic explanations are intrinsic to the various interests we have in our relationships with each other. In the introduction I noted that it would not be sufficient for the defence of my thesis simply to point to the fact that these interests are different from the interests served by physicalistic explanations, and then conclude on that basis that the work carried out by mentalistic and physicalistic explanations must therefore be completely separate. Unfortunately, the fact that different forms of explanation answer to different types of interest does not yet guarantee that the work carried out by the one is completely separate from the work carried out by the other. For that reason, I said that it would be necessary to directly attack the idea that mentalistic and physicalistic explanation must converge on a

common subject matter. But in light of my argument concerning the intrinsic connection between the nature of the mental and the nature of human relationships, I now want to close this chapter by suggesting a reason for thinking that these two stages are logically connected, which has so far remained implicit in the background.

What I want to suggest is this: if the nature of the interests served by mentalistic explanations are tied to the nature of our everyday relationships with each other, then the question of whether the work carried out by mentalistic explanations is completely separate from the work carried out by physicalistic explanations can be settled by determining whether the nature of our mentalistic interests merges in some sense with the nature of our physicalistic interests, which in turn can be settled by determining whether our everyday relationships are motivated out of a concern with the prediction and causal explanation of each other's behaviour. What I am effectively claiming here is that if the nature of mentalistic interests is tied to the way in which individuals actually relate to each other on an everyday basis, and if we actually relate to each other with a view to predicting and causally explaining the other person's behaviour, then it must seem natural to hold the view that mentalistic explanation is a form of causal explanation; and hence, it must seem natural for the non-reductive physicalist in particular to hold that mentalistic explanation is a form of causal explanation which derives its explanatory efficacy from that of physicalistic explanation. But on the other hand, if our everyday relationships are not motivated or driven by such concerns, and if we do not commit ourselves to the same metaphysical principles, then recognition of the autonomous nature of mentalistic explanation in the deeper sense is arguably the next logical step to take.

Chapter 4: Thinking and Relating

1.Introduction

In this chapter a more precise account will now be given of the nature of human relationships, and its implications for the autonomous nature of mentalistic explanation will be brought to light. This will be achieved by developing features of MacMurray's account of the nature of human relationships in the direction suggested by my interpretation of Wittgenstein's rule following considerations in the previous chapter. The result of this combination will be an account of the nature of human relationships and the nature of the mental which will contrast directly with the account of human relationships to which the physicalist's conception of the mental is implicitly committed. More specifically, on the basis of this combination it will be argued that there is an intrinsic connection between what it means to say that an individual has the capacity to think and what it means to say that he has the capacity to be involved in various types of human relationships. This will provide a concrete example to show how the nature of the mental is tied to the nature of our relationships with each other, and it will also show that the extent to which the autonomous nature of mentalistic explanation can be acknowledged is tied to our account of the way in which we actually relate to each other on an everyday basis.

2.More on human relationships

2.1.A Four-Fold Division

It is by no means obvious that there is an intrinsic connection between what it means to say that an individual has the capacity to think and what it means to say that he has the capacity to be involved in human relationships. It does not seem necessary to having the capacity to think that an individual has the capacity to be involved in relationships with others, nor does it even seem necessary that other people are there in the first place. So there is bound to be a strong tendency to deny that thinking and relating are intrinsically connected, but it seems to me that there are nonetheless some strong arguments in favour of this claim. The intrinsicness of this connection can be brought into sharper focus by concentrating on a particular way of dividing up the general category of human relationships, which will indicate some definite points at

which having the capacity to think can be seen to presuppose having the capacity to be involved in relationships with others. The general category of human relationships is too wide and unspecific as it stands, so I will make use of the four-fold division of human relationships into personal and impersonal human relationships, and direct and indirect human relationships, which is suggested by MacMurray:

> our relation to another person may either be personal or impersonal. Like all our relations to the Other, these are primarily practical, but they have, of course, their theoretical aspect. Each, therefore, gives rise to a knowledge of people. The first gives rise to that personal understanding of others which is the result of reflection upon our personal dealings with men and women of varying sorts under varied conditions, and which we sometimes call 'a knowledge of the world'; the second, if it is systematically pursued, leads to the scientific knowledge of man and his behaviour (1970: 30)

> We must distinguish between the direct and indirect relations of persons...Direct relations are those which involve a personal acquaintance with one another on the part of the persons related. Indirect relations exclude this condition: they are relations between persons who are not personally known to one another. All indirect relations are therefore necessarily impersonal. Direct relations are those which may or may not be personal, at the will of the persons related. If they are maintained at an impersonal level, this requires a justification. (1970: 43).

Personal human relationships are relationships in which the individual's involvement is spontaneous and immediate in everyday life; they are relationships between persons in their capacity as friends, lovers, work colleagues, and so on, and as such these relationships are characterised by various types of demands, which can only be described using the moralistic and intentionalistic language particular to this domain of human life. Impersonal human relationships, on the other hand, are relationships in which the individual's involvement is rather limited to certain contexts, such as relationships between scientists and their subjects in the context of the laboratory, where the purpose of entering into these relationships might be to discover the effects of certain types of drugs on the brain and the nervous system, for example. These particular impersonal relationships are characterised by demands which can only be described using the explanatory resources particular to that field of scientific investigation.

Human relationships are further divided into direct and indirect relationships. Direct relationships are relationships in which the individual's involvement is of immediate concern to him, with someone with whom he is personally acquainted, such as his involvement with his loved ones and friends. Indirect relationships are those in which the individual's involvement is of less immediate concern to him, unless they are disrupted for some reason, and in which his involvement is with people with whom he is not personally acquainted, such as his relationships with the supplier of his furniture and the baker of his bread. Although the

latter are also impersonal, they differ from the type of impersonal relationship mentioned a moment ago. We do not personally know the baker or the furniture supplier, but we do not thereby relate to them with a view to acquiring 'scientific knowledge of their behaviour'.[1]

The division of human relationships into personal and impersonal relationships does not correspond exactly to the division into direct and indirect relationships, although it is closely connected to it. Direct relationships might be either personal or impersonal. Direct relationships are personal when they are carried out on an ordinary everyday basis in a variety of situations with people with whom we are personally acquainted. Yet there might be specific occasions on which direct relationships are carried out on an impersonal level- when an individual studies the behaviour of his friend in a controlled situation in order to predict how he will be affected when his body is deprived of nicotine. Indirect relationships, on the other hand, are necessarily impersonal, in the sense that they involve people who are not personally acquainted, even though they are not necessarily carried out with a view to gaining scientific knowledge of the other. So a relationship is direct if it is between persons who know each other, otherwise it is indirect and between persons who do not know each other. A direct relationship is personal if it is carried out on an everyday level between persons who know each other, and it is impersonal if it is carried out on a level which is more appropriate to gaining scientific knowledge of the other. An indirect relationship is necessarily impersonal, but it may be impersonal in a straight-forward everyday sense that it is carried out between persons who are not personally acquainted, as in the case of the baker and the furniture supplier, or it may be impersonal in a non-everyday sense that it is carried out between persons who are not personally acquainted, but where the aim is to gain scientific knowledge of the other. Macmurray writes:

> The personal relation with the other is possible only between persons who know one another. But our own personal activities depend upon the personal activities of large numbers of people whom we do not and cannot know. All my activities have an economic aspect, for example. I need food; consequently I depend upon a host of people who produce, transport and deliver food to me. When I pay for food; I contribute my quota of assistance to the personal lives of all these people. One aspect of my dependence is my belief that their personal activities will continue in the future as they have done in

---

[1] MacMurray uses the term 'impersonal' to characterise direct relationships entered into for the purpose of gaining scientific knowledge of the other, and to characterise all types of indirect relationships. For the purposes of clarity, I will find it necessary at certain points in my argument to mark this distinction by using the higher-level categories of 'everyday relationships' and 'non-everyday relationships'. I will therefore stipulate that 'everyday relationships' include both direct personal relationships, and indirect impersonal relationships carried out on an ordinary basis (baker and furniture supplier); and I will stipulate that 'non-everyday relationships' include both direct impersonal relationships, and indirect impersonal relationships carried out with a view to gaining scientific knowledge of the other.

the past. I must trust in the continuance of patterns of activity carried on by persons whom I do not and cannot know. The relation so established between myself and them is a relation of persons. But the relation is necessarily impersonal; and consequently the knowledge on which it rests must be merely objective. I must conceive the activities of those others upon whom I depend as automatic and continuant, although I know well enough that they are personal doings. (1970: 43).

What this suggests is that people with whom an individual is involved in indirect relationships are interchangeable in a way in which people with whom an individual is involved in direct relationships are not. It is of much less concern who is there to supply the furniture and bake the bread than who is there to listen to our good news and take pleasure in our happiness. It is of much less concern who is there to process our bank cheques and drive our trains than who is there to offer support and encouragement in difficult times, or to share a walk, a meal or a film. The most basic and immediate of human relationships in which an individual is involved are therefore direct personal relationships, which emanate outward to encompass the various indirect and impersonal relationships that make up the vast web of human relationships in which the individual is involved, to greater or lesser degrees. Indirect relationships are thus more distant from the individual's concerns and interests than are his direct relationships, and this means that his attitudes to those people with whom he is involved in direct relationships will differ significantly from his attitudes to those people with whom he is involved in indirect relationships.[2]

This point has to be treated with caution, due to its importance: corresponding to the wider category of human relationships is the wider category of attitude to persons, whilst corresponding to the division of human relationships into direct and indirect relationships is the division of attitude to persons into direct and indirect attitudes to persons. An individual's attitudes to the supplier of his furniture or the baker of his bread is indirect in that his relationships with these people do not involve any form of emotional or affective attachment, whereas his attitudes to his family and friends are direct in that these forms of attachment are partly constitutive of his involvement in such direct relationships. An individual's attitudes towards other people not only manifests themselves in his treatment of them as rational agents in their own right, but also in his being able to respond to the demands which are constitutive of the relationships in which he is involved. This applies to the individual's direct and indirect attitudes alike, differing only in the types of demands to which his attitudes alert him. Without having the capacity to be emotionally and affectively attached to others, the individual would be unresponsive to the demands which are constitutive of relationships

[2] In what follows I will go beyond what MacMurray explicitly says.

between loved ones, friends and family members; without having the capacity to be economically and legally committed to others, the individual would be unresponsive to the demands which are constitutive of relationships between traders and suppliers of various types of goods and services.

2.2a.Attitude To Persons As A Basic Orientation

Direct and indirect attitudes to persons are specific manifestations of our attitude to persons in general. Whereas direct and indirect attitudes to persons are open to moral appraisal and assessment, it does not seem correct to raise the question of the moral praiseworthiness or blameworthiness of our attitude to persons in its broadest form. Our attitude to persons in its broadest form cannot be regarded as being based on judgement or deliberation, or as being demanded of those individuals who are committed to the types of responses consititutive of a particular moral outlook. To say that our attitude to persons in its broadest form is not open to moral appraisal, in the way that its more specific manifestations are, is simply to say that there is no genuine question of its being appropriate or inappropriate with respect to that toward which it is an attitude. I think that part of the reason for this is that our attitude to persons in its broadest form is spontaneously solicited from the individual in such a way that it is already presupposed as being in place, before any question can arise with respect to the appropriateness of its more specific manifestations within direct and indirect human relationships. It might be helpful, as a way of expressing this spontaneity, to say that our attitude to persons in its broadest form is a basic orientation that we have toward other living things, which differs in significant ways from the basic orientation that we have toward non-living things. What is distinctive of living things is that they have the possibilities of movement and expression which command a certain type of attention and respect not commanded by non-living things, and it might be said that what is distinctive of persons in particular is that they command the type of attention and respect that immediately draws us into acknowledgeing their *presence* in a situation *with us*.

Particularly relevant in drawing us into acknowledging the presence of other people is the fact that their possibilities of movement and expression are limited within the human bodily form. The limitation here is not to be understood negatively; the limits of the human bodily form do not seem to be circumscribable with any degree of precision or mathematical accuracy. If the limits of the human bodily form were circumscribable at all, I suspect that the specification of

these limits would have very little to do with our acknowledgement of a person's presence in a situation with us. What I think is important is that since the movement and expression of other people is what is responsible for drawing us into an acknowledgement of their presence, the human bodily form must figure quite centrally in an explanation of what it means to have the capacity to be involved in various types of human relationships. This point can also be put by saying that having the human bodily form is a logical precondition of being able to have, and express, the array of feelings and emotions which are partly constitutive of many of our relationships. This way of putting the point is suggested by McClintock, who illustrates the importance of the human bodily form in our relationships with reference to certain types of feelings and emotions which are only expressible through touch, and other more subtle forms of bodily contact:

> The kinds of feelings and emotions...which can find expression through touch are conceptually connected to, amongst other things, the following physiological details: Having arms and legs, having hands which can bend in such a fashion as to be able to follow the contours of a body, having a body that varies in sensitivity with mood...having a bodily sense that can be aware of limb position or muscle state without observation, having a body such that tension can lead to pains in the neck or a racing heart, having a body that is capable of shaking with fear and of sweating with anxiety and so on. (1995: 90).

The point that McClintock is making here is that there is a logical connection between certain types of feelings and emotions that an individual is capable of having and the bodily capacities and attributes that give expression to these feelings and emotions. It might be said that attributing emotions or feelings to individuals is in most cases a matter of being involved in a direct relationship with them, and it might also be said that a logical precondition of this is that the individuals involved have the human bodily form. What explains this point is that the human bodily form commands our attention in such a way as to spontaneously solicit the attitude toward persons that sustains the different types of relationships in which we can be involved. If this is correct, then the possibility of being involved in relationships with others depends on at least these two basic factors: first, it depends on our having the basic orientation toward others which is an acknowledgement of their presence in situations with us; second, it depends on our both having the human bodily form, which gives expression to the basic orientation in one of us, and to which that basic orientation is logically tied in the other, and *vice-versa*. Having this basic orientation is therefore a logical precondition of having the more specific direct and indirect attitudes, which are necessary if we are to have an awareness of the demands which other people impose on us. With regard to the more specific

manifestations of our basic attitude to persons, the question of their appropriateness can be raised.

2.2b.The Appropriateness Of Direct And Indirect Attitudes

The various direct and indirect attitudes that an individual has toward others sustain his relationships with them. Direct and indirect relationships are constituted by demands which must be acknowledged if the individual is to be said to have the capacity to be involved in them. But it seems to me that an individual can only be said to have an awareness of the demands of his relationships if he has the appropriate attitudes towards those people with whom he is involved. Direct and indirect relationships are constituted by different types of demands, so individuals who are involved in such relationships must have different types of attitudes to different people, depending on the nature of the relationship in question. Direct and indirect attitudes to other persons can therefore be judged to be appropriate relative to the relationship in which they are involved, since it is only in terms of these relationships that the individual's attitudes can be considered to be open to moral appraisal. The individual can be said to have the capacity to be involved in various types of relationships in so far as he can be said to have an awareness of the demands that these relationships present to him, and he can be said to have an awareness of these demands in so far as he can be said to have the appropriate direct or indirect attitudes to other persons.

Cockburn (1990) suggests that we come to think of others as beings toward which certain attitudes and responses are appropriate as a result of our training, which instils in us the propensity to give, and accept, reasons for feeling certain things, and acting in certain ways. Cockburn's idea seems to be that the child in his pre-linguistic state is naturally disposed to respond in various ways to other individuals with whom it comes into contact, but that it hardly makes sense to say that the child's responses at this stage of its development are manifestations of the fact that it thinks of other individuals as beings toward which certain responses are appropriate or inappropriate. The child might naturally respond caringly to his sister when she falls over, but he does not yet think of his sister as an individual toward whom a caring attitude is appropriate. This can be seen in the fact that the child might rather hit or kick his sister whenever she happens to frustrate his current pursuits, and her falling over on this occasion might have done just that. Until the child has acquired the linguistic skills and abilities which are integral to the way of living into which he is being brought up, it

makes no sense to say that he thinks of other individuals as making any kind of demands on him. It is only through being brought up into a particular way of living his life with others that he acquires the linguistic skills necessary for him to think of them as beings toward which certain responses and attitudes are appropriate. Cockburn notes that:

> the possibility of saying, in any rich sense, that it thinks of others as beings who are to be treated in certain ways only emerges gradually as the child develops. It cannot be said that the very young thinks of the situation as presenting him with reasons for feeling and doing certain things. A clear foothold for that way of speaking only emerges when he begins to give and accept reasons for feeling and acting (1990: 7).

I think that this goes some way toward helping us understand what it means to talk about the appropriateness of our attitudes in our relationships with others. It brings out the importance of linguistic training in the development of our attitudes to others which, as a result, issue in responses to individuals whom we think of as persons who ought to be treated in certain ways. Without having the relevant concepts, the individual would not have the capacity to think of others as persons who ought to be pitied, persons who ought to be respected, persons who ought to be feared, and so on. It is only once the relevant conceptual skills are acquired that it begins to make sense to talk of our attitudes to persons as developing in more specific ways within the contexts of different types of relationships, for it is only once the relevant skills are acquired that it makes sense to talk of the individual as being aware of the demands that other people in his relationships present to him. But this takes us to a deeper point concerning the appropriateness of our attitudes, which expands on the idea that our attitudes to others are developed in more specific directions through the acquisition of conceptual skills.

Winch (1987) argues that part of what it means to say that our attitudes and responses to other individuals are appropriate is that on the particular occasion on which they are expressed, they can be said to be instances of more general attitudes and responses, which would be expressed on other occasions toward other people in similar circumstances. Winch refers to this condition for the appropriateness of our attitudes as the generality condition, which effectively requires that an individual's reaction to a particular person on a particular occasion be typical of that individual's reaction to any other person on any other occasion. Of course, this must be complicated by the fact that an individual might not always react as he really ought to react, even if there had been no question that he had reacted appropriately to another person on another occasion. An individual who reacts sympathetically to his friend

when she is distressed at the death of her mother might not react in this way to another friend in similar circumstances, and his failure to react in the same way would require some sort of explanation. Perhaps the individual was particularly fond of his friend's mother himself, so much so that he found it hard to cope with her death or offer comfort to his friend in the way that he would normally have been expected to. At any rate, allowing for such complications, an individual's attitudes to others are typical of the attitudes that any individual who is involved in such direct relationships would be expected to have in these circumstances. And since it is perhaps too much to expect the same attitude from an individual who is not directly involved in this relationship, the generality condition can be read as being sensitive to the variations in individuals' attitudes across different types of direct and indirect relationships.

Winch's deeper point is that the generality in our responses cannot be understood independently of the fact that they are regular and constantly repeated features of human life, which means that it is only through learning how to successfully participate in on-going relationships that an individual acquires the concepts necessary to appreciate the generality in the particular response. It is important to stress that the individual comes to appreciate the generality in the particular response through acquiring the conceptual skills which are exercised in the expression of his attitudes to those persons with whom he is involved. It would be wrong to think that the process of acquiring the relevant conceptual skills could proceed independently of the process of learning how to relate to others, so it would be equally wrong to think that an individual's capacity to exercise his conceptual skills in the expression of his attitudes could be explained independently of explaining his capacity to be involved in relationships with other people.

It seems to me that this deeper point helps us to understand the following connection: in learning to use the concepts 'suffering' and 'distress' in relating to his friend, the individual is at the same time learning to use the concepts 'comfort' and 'pity' to express his own attitudes to his friend. Learning to use these concepts together is a matter of learning how to be involved in relationships with other people: it is a matter of learning how to say the right word in the right tone of voice; it is a matter learning how to offer emotional and practical support; it is a matter of learning how to look into her eyes and calmly talking to her to put her at ease, and so on. Through being involved in situations like this one, the individual experiences different aspects of different types of relationships, and he learns how to see and cope with the demands of the situation as they arise within the context of these relationships.

The upshot of this is that the individual learns to conceptualise 'suffering' together with 'pity', 'distress' together with 'comfort', and he comes to think of an individual who is suffering, or an individual who is in distress, as a person toward whom pity and comfort are appropriate attitudes to have. The appropriateness of the individual's attitudes to his friend in this particular direct relationship is therefore decided relative to the demands presented to him in it. The friend's distress and suffering at her mother's death presents the individual with a reason for comforting her and offering her pity, and part of what makes the individual's responses appropriate on this occasion is that the conceptual skills which are exercised in the expression of his attitudes forge an internal connection between his attitudes and their object, and this means that the attitudes expressed on this occasion can be expected to be in place with respect to other people on other occasions.

Winch's point is that the conceptual skills exercised in the expression of the individual's attitudes of pity and comfort tie the suffering and the distress intrinsically to the attitudes themselves, such that the individual's attitudes of pity and comfort take the other person's suffering and distress as their proper objects on this and on other occasions. But again his point has to be put quite carefully, since there are difficult cases which do not seem to conform neatly to these conditions. It might be that an individual responds caringly to one person's suffering, yet indifferently to another person's suffering, more so when that person is not personally known to the individual; or in an extreme case, it might be that the individual is incapable of displaying any emotional response whatsoever to the other person's suffering, perhaps through having had an atrocious upbringing, or perhaps through being psychologically dysfunctional, preventing the individual from living a normal life in the company of others.

But what these extreme cases indicate is not that suffering cannot be deemed the proper object of the attitude of care, or that distress cannot be deemed the proper object of the attitude of comfort, but that these individuals are, for some reason or other, incapable of having what we would be inclined to call the appropriate responses on these and perhaps on other occasions. What constitutes an appropriate response in such cases is certainly not susceptible of strict definition, but there are various factors which we would normally expect to be in place as the norm. We are brought up to see that other people make demands on us, and that we ought to have certain attitudes and responses in certain situations. There will inevitably be cases in which certain individuals stand out as exceptions to the norm, but I think it would be erroneous to take these individuals as the norm rather than as exceptions. To

suppose that such individuals were not exceptional cases would be to countenance the possibility of a complete break down in the stability and continuity in our human relationships, in our explanatory practices, and in our everyday lives in general, threatening the very existence of what we call the norm in these situations.

3. The intrinsic connection between thinking and relating

3.1a. Responding To The Demands Of Relationships

I now want to explain what it means to say that there is an intrinsic connection between what it means to have the capacity to think and what it means to have the capacity to be involved in various types of human relationships. What this amounts to is that there are certain situations in which an individual can be said to be thinking if he can be said to be responding to the demands of the situation in which he finds himself, and that his awareness of these demands is inseparable from his awareness of the demands of the various human relationships in which he has the capacity to be involved. Or in other words, what it amounts to is that there is an intrinsic connection between the individual's awareness of the demands that the situation presents to him, and his awareness of the demands of the direct and indirect relationships in which he happens to be involved. My point is not that every case of thinking is a case of responding to the demands of the situation in which the individual finds himself; nor is it that every case of thinking is a case of being involved in a particular direct or indirect relationship. That would be an extremely difficult position to maintain, and I am sure that it would not be correct. Individuals can be lost in private thoughts, just as they can be responding to the demands of their situations; and solitary individuals can still be said to have the capacity to think, despite having withdrawn themselves as far as possible from all contact with other people. My point is rather that an individual can be said to have the capacity to think in so far as he can be said to have the capacity to be involved in human relationships, and solitary individuals who have withdrawn themselves from contact with other people can still be said to have this capacity.

The idea can be illustrated if we reconsider the case of the individual who is moved to comforting his friend on the death of her mother. Within the context of this particular direct personal relationship, the individual's involvement is such that he is responsive to the distress and suffering that her mother's death has caused her. Her suffering presents itself to the

individual as a reason for consoling her, and in responding to his friend in this way he can be said to be responding to the demands of this relationship. At the same time, it can be said that it is his involvement in this relationship that alerts him to the demands of the situation in which he finds himself, in which his friend is not eating or sleeping properly, in which she is finding it hard to get on with her normal routine, in which she is spending most of her time in solitude, and so on. The individual responds to the demands of this situation by comforting his friend and encouraging her to return to her normal routine, and in so doing it can be said that he is responding appropriately to the demands which are presented to him in this relationship. The individual's awareness of the demands of the situation is shaped by the conceptual skills which he has acquired in the process of learning how to be involved in direct personal relationships with others, and his awareness of these demands is therefore intrinsically connected to his awareness of the demands of this direct personal relationship in which he is involved with his friend. So it can be said that the individual thinks that his friend is distressed, that she is suffering, and that she ought to be comforted, in so far as it can be said that he is responding as he does to the demands of the situation to which his involvement in this relationship have alerted him.

The same point can be illustrated with the case of an individual's involvement in an indirect relationship, but this time I will mention the types of demands which are often overlooked. Suppose that the individual is a carpenter who is busy at work on a dining set that he has been commissioned to make. He knows that the dining set must be made according to the customer's specifications, otherwise he will fail to meet the demands of the contract. That much is obvious. But he also knows that the chairs must be built in proportion to the table, that the chairs must not be so light as to collapse the minute they are sat on; he knows that the chairs must not be so heavy that the guests cannot adjust their position relative to the table, and he knows that the chairs must be wide enough that the guests can sit down comfortably without falling off; he knows that the table must not be so low that the guests would have to sit on the floor to eat from it, and he knows that there must be sufficient space around the table for the guests to sit without rubbing shoulders and bumping elbows. That much is also obvious. So obvious, in fact, that there is nothing in the contract to cover these demands. They are simply taken for granted as constitutive of the practice of sitting down at a table to share a meal.

Suppose that in building one of the chairs the individual stops what he is doing for a moment, runs his hand over the seat of the chair, and then reaches for the sandpaper that is lying next to him on the workbench; or suppose that he stops what he is doing, carefully studies the lengths of the legs, and then gets the plane that is hanging on the wall beside his bench. Can we say that the individual is thinking in these cases, and that in so doing he is responding to the demands of some relationship? It seems to me that we can, for we can say that in his work he is responding to the demands of the situation as they arise for him, and we can say that his reaching for the sandpaper is expressive of his awareness of the fact that the seat of the chair is still too rough for anyone to sit on it comfortably without getting a splinter, and we can say that his getting the plane is expressive of his awareness of the fact that the chair will be too wobbly for anyone to sit on it if one of the legs is longer than the other three. So it can be said that the individual is thinking in so far as it can be said that in his work he is displaying an awareness of the demands of the situation as they arise for him, and that his awareness of these demands is at the same time inseparable from his awareness of the demands of the indirect relationship in which he is involved in building this particular dining set for his customer.

There is one final illustration of this point that I want to consider, which places greater emphasis on the individual's involvement in both direct and indirect relationships at the same time.[3] An individual is at confession, relating a particular venial sin to the priest, when he pauses and stares off into space for a moment; an aspect of the sin suddenly strikes the individual which he had not considered before, and he comes to realise that because of the negative way in which his sin touched the lives of other people, his sin was in fact a mortal sin. The individual's thinking in this situation is embedded in the traditions and the doctrines of the church, and in the deeply religious way of life into which he has been brought up. His awareness of that aspect of his sin which renders it a mortal sin, as opposed to a venial sin, is inseparable from his involvement in various direct and indirect relationships, including his direct personal relationship with the priest, his direct personal and indirect impersonal (in the everyday sense) relationships with the larger body of people who share this religious way of life, and who uphold the traditions and doctrines which qualify the individual's sin in this way, and his direct (and possibly indirect) relationships with the people whose lives were negatively affected by his sin. The individual's awareness of the sin as a mortal sin is therefore inseparable from his awareness of the demands of the various direct and indirect

[3] This particular example is based on an example from Canfield (1994).

relationships in which he is involved, in carrying on and upholding the traditions of the religious way of living into which he has been brought up.

In the different cases that I have considered, the individual's thinking has been said to be intrinsically connected to his responding to the demands of the various direct and indirect relationships in which he is involved. The individual's responses to these demands are open to moral appraisal and rational assessment in terms of the demands themselves, an important consideration in which is that the individual has the appropriate direct or indirect attitude to whomever he is involved with in his relationships. The final point I want to make in this connection is one which presents an expansion of Wittgenstein's idea, that the concept thinking comprises within itself many manifestations of life (1990, § 110). It is not immediately obvious what Wittgenstein is getting at here, but it seems to me that it is not unreasonable to interpret it within the context of this discussion, as stating that the individual's responding to the demands of his relationships can be said to take a variety of different forms, and which form is taken can depend on many factors. It can sometimes depend on the individual's level of skill and competence, other times it can depend on the individual's personal mood and state of mind, yet other times it can depend on the type of relationship in question. The individual who thinks that his friend ought to be comforted responds caringly toward her, and his loving attention can be said to be the most appropriate form that his awareness of the demands of this situation ought to take. His involvement in this direct personal relationship is such that his awareness of the demands is immediate, in that there was no need for a process of deliberation over whether or not his friend actually was distressed.

The individual who is confessing his sins to the priest suddenly realises that his sin was a mortal sin, and his feelings of deep sadness and regret over his actions is the form that his awareness of the demands happen to take. His awareness of this aspect of his sin might have further manifestations, perhaps in his denying himself some of his normal pleasures, or in his subsequent efforts to make amends for the hurt that he has caused others. In this case the individual's awareness of these demands struck him only after discussing the problem over with the priest, but it might have been that the individual had privately come to this realisation himself through considering the motive behind his action, and then coming to the decision that he could not reconcile his motivation with the fact that his action had caused so much hurt, and then on the basis of this he might have finally arrived at the conclusion that

his action would have to be seen by the church in a certain light. The important point to note here is that the individual can be said to be thinking in certain situations where he can be said to be responding to the demands of these situations as they present themselves, and that his awareness of these demands, which is intrinsically connected to his awareness of the demands of the direct and indirect relationships in which he happens to be involved, can take many different forms. [4]

3.1b. What About The Solitary Individual?

The obvious objection to make against this approach to thinking is that it places too much emphasis on the individual's involvement in relationships with others. An individual can surely cut himself off from others and commit himself to a life of solitude and isolation, in which his activities and routines can no longer be said to be part of his involvement in any relationship with other people. Within that state of withdrawal the individual can be said to act for the reasons that his situation presents him with, reasons which might only be intelligible from within the context of his life in isolation from others. Or at least, it can be said that he can act for reasons which are not intrinsically connected to the demands of any type of direct or indirect relationship. He might think it prudent to reinforce his shelter when he notices the darkening clouds, for instance, or he might think it fortuitous that the wood is still dry enough to light his fire. But far from presenting counter-examples to the position I have been developing, it seems to me that they can be dealt with in a satisfactory manner.

---

[4] There has been some amount of debate recently on whether Wittgenstein held that thinking can sometimes be said to be a mental process or activity. The debate has largely been a reaction to Hacker (1993), who argued that thinking cannot properly be said to be a process or an activity because it lacks the grammatical features of a 'typical' process or activity, the most central of which seem to be that thinking does not have genuine duration, and it does not have the right kind of structure. The same points are echoed by Schroeder (1995), who argues that thinking is characterised by a temporal indeterminacy which does not characterise typical processes and activities, and that thinking lacks the exact micro-structure which typical process and activities possess. On the other side of the debate, Hanfling (1993) argues that there are activities and processes which thinking can sometimes be said to consist in: thinking can be said to take the form of an activity when a musician is working out a piece of music at the piano. His composing at the piano can be said to be what his thinking consists in. Thinking can also be said to take the form of a process, when an individual is mulling over a particular problem by weighing up alternatives and deciding to accept or reject certain ideas. His speaking can be said to be what his thinking consists in. It seems to me that Hanfling is on the right lines. Hacker and Schroeder have concluded that thinking cannot be a process or an activity because it lacks certain features of 'typical' processes and activities; but it is not clear that there is a 'typical' process or activity to which thinking must conform if it is to count as a ***mental*** process or activity. That thinking sometimes takes the form of a process or activity cannot be rejected on the grounds that it lacks the grammatical features of typical ***non-mental*** processes and activities. See also n.5.

Part of what it means to say that an individual has the capacity to be involved in human relationships is that he has the capacity to withdraw from his involvement with others and to be alone. This is essentially Heidegger's idea, only expressed in different terms, that being-alone is a necessary determination of being-with others (1967: 156-7). It is also part of what it means to say that an individual has the capacity to be involved in human relationships that he has his own individuality and personal space, which enables him to act for reasons that are presented to him within a situation that can be his own making. Having the capacity to be involved in human relationships empowers the individual to withdraw from his involvement and take up a way of life in which his contact with others is perhaps non-existent, and it empowers the individual to exercise his conceptual skills on some occasions with such a degree of freedom and privacy, that it *seems* to allow us to dispense with his involvement with others as a necessary precondition for saying that he has the capacity to think in the first place. But it seems to me that the capacity for freedom and privacy, for individuality and separateness, are simply aspects of what it means to have the capacity to be involved in human relationships. So it is only to be expected that an individual who has the capacity to be involved in human relationships, and hence to think, is an individual who has the capacity to lose himself in his own private world, or to live out a solitary existence if he so desires.

3.2. Uses Of The Concept Thinking As Acts Within Human Relationships

At this point I want to develop certain aspects of Wittgenstein's recommendation that we look at the word "to think" as a tool which is used within the contexts of our life with others (1967: § 360). I am not so much interested in developing the tool metaphor, rather than the suggestion that the word "to think" can be used for various purposes within the context of various types of human relationships. Wittgenstein's recommendation to consider the different uses that we actually make of our psychological concepts seems to be in danger of being dismissed out of hand, as an unsuccessful attempt to avoid metaphysical problems by an irrelevant appeal to what we say in everyday life. But this is a negative way of taking Wittgenstein's recommendation, which might be better understood as a recommendation to look at the different ways in which the concept thinking is used as part of the way in which we relate to each other. This is positive, and more effort is required in its execution than it seems. But this positive understanding of Wittgenstein lets us say that certain uses of the concept thinking are acts which are partly constitutive of some of our relationships, acts of the type which create commitments and obligations for the individuals who are involved in

them. More than this, it lets us say that there are certain uses of the concept thinking which are logically connected to the uses of a host of moral and inter-personal concepts, such as 'commitment', 'obligation', 'duty', 'agency', 'personhood', and so on.

Hunter (1990: 59ff) brings out the connection between certain uses of the concept thinking and uses of the concepts of obligation and commitment, by considering cases in which an individual uses the concept thinking (in a direct relationship) to indicate non-committal to a suggested course of action.[5] The connection is brought out most clearly when the individual avoids committing himself to acting in a certain way by saying that he will have to think about it first. It is characteristic of this use that although the individual is indicating that he is not prepared to commit himself to the course of action at that moment, he is nonetheless seriously inclined to adopt this course of action at some stage or another. If it later came out that he had no such inclination, we might be entitled to conclude that he had intended to deceive us or that he did not want to disappoint us by declining the offer out-right. Suppose, for example, that an employee is offered the opportunity to manage the office for a week whilst the boss is on a business trip, and he says that he will think about it. The boss is then entitled to expect that the employee is seriously inclined to take up this opportunity, otherwise he would not have indicated his willingness to consider the offer by saying that he would think about it.

But suppose that the employee is actually quite frightened by the prospect of being in charge for a week, and he really believes that he would simply be unable to cope with the responsibilities. The problem is that he has created certain expectations by indicating a willingness to consider the offer, when in fact he had no serious inclination to take it. In this case, saying that he would think about it would not have been an indication of his willingness,

---

[5] Hunter (1987: 118-120) discusses further uses of the word 'think', concerning which I am in broad agreement. However, I am not convinced by the central claim of that article, that in some cases, where the word 'think' is used to scold someone's substandard performance, as when we say 'you weren't thinking', we are *pretending* that an auxilliary activity of thinking should have been engaged in to bring about the best results, when in fact it wasn't. Hunter argues that: "As an indirect way of saying this, we trade on the conception of thinking as an auxilliary activity enabling us to perform certain tasks competently...we pretend that the beneficial activity did not occur, as a way of saying that the effect it might have had did not occur...but all we are actually saying is the latter." (126-127). It seems to me that such pretence is no part at all of making the latter claim, and I think that Scheer (1991) is correct to say that the auxilliary activities not performed, for which the person is scolded, are those of pausing, checking, comparing, etc; so in saying 'you weren't thinking', we are simply referring to the fact that *these* auxilliary activities were not performed. We are *not* pretending that there is a further auxilliary activity, an imaginary one, underlying the person's overt actvities that would have brought about the appropriate standard of work.

as he led others to believe; it would rather have been a way of avoiding the embarrassing situation of having to admit to his boss that he was not ready for the promotion. This use of the concept thinking presupposes that the individual is capable of being involved in certain on-going personal relationships, and it also presupposes that he has an awareness of the demands of that aspect of his relationship with his boss, which he had created by saying that he would think about the offer. It is an act which is itself partly constitutive of this relationship, dictating the manner in which it really ought to evolve. It has added a new dimension to the relationship, and with it a new set of demands which, in this case, the individual openly acknowledged but privately tried to avoid.

There is yet a further use of the concept thinking which can also be said to commit the individual to a certain standard of behaviour, and to a certain constancy in his judgements, when it is used to express an opinion, or to make a judgement. An individual can take a moral stance by saying that he thinks that brutality to animals is a crime, or that he thinks that children should be disciplined from an early age. Interestingly, this use of the concept thinking is more or less interchangeable with certain uses of the concept believing, in that an individual who thinks that brutality to animals is a crime is an individual who believes that brutality to animals is a crime, and an individual who thinks that children should be disciplined from an early age is an individual who believes that children should be disciplined from an early age. But perhaps a more important point to note here is that in using the concept thinking in these ways the individual is expressing a persisting moral attitude, raising expectations in others as to how he would be likely to act in certain situations, or as to what he is likely to think about some other issues. The concept thinking can also be used to express an opinion or to make a judgement when the individual is not entirely sure of his grounds, and that he is quite prepared to retract his statement if he finds any good reason to do so. Although this use of the concept thinking can again be said to create expectations in others as to how the individual is likely to act, the expectations must be weakened by the fact that in using the concept thinking in this way the individual is indicating that he is not prepared to meet the demands that he has created, come what may. The statement, 'I think he is coming today, but I am not sure', indicates that the individual is not prepared to defend this claim to the last, and that it should not be unexpected if he were to retract it.

Even if the individual does feel certain that he is correct, in which case he would not have said, 'I think...', but simply, 'He is coming today', he can later qualify this statement by using

the concept thinking to justify his preparing tea for a visitor, when the chances of his coming seem to be lessening: 'Well, at least I *think* he is coming today!' This use of the concept thinking is in a sense justificatory, presupposing that the individual is fully aware of the expectations which he has created in others when he made his original statement. If he were not aware of the expectations that his statement created, he would not have felt it necessary to justify his making this statement by using the concept thinking when it started to look as if the expectations would be frustrated. This same use can be illustrated in the case where we try to make sense of an individual's misguided actions, as when we say, 'She thinks her keys are in the drawer, but I know they are not'. If she were asked why she was rummaging in the drawer when her keys were actually on the table in the next room, she might reply, 'well, I thought they were in there'. The important point to note here is that an individual who uses the concept thinking in this justificatory sense must not continue to search in the drawer after she has been told that her keys are elsewhere, and after she justifies her mistake by saying that she had thought her keys were in the drawer. It would be strange for her to continue searching in the drawer after justifying her actions in this way, but if she were to continue searching in the drawer nonetheless, we might be entitled to conclude that she still strongly believed that her keys were in there, and that she suspected that others were trying to mislead her in telling her that the keys were on the table in the next room.

4.Interpretationism

4.1 The Interpretationist's Account Of Thought

It might seem at first glance that the account of thought I have been developing is not too different from the account of thought developed by the physicalist, and that the approach to the autonomous nature of mentalistic explanation that I want to recommend, as a consequence, will not be much different either. The need to regard human interaction as the basis for an account of thought can be seen to be integral to the physicalist's position, just as much as mine, as stemming from the need to create a suitable context in which the notion of rationality can be applied to human beings. On the assumption that human beings are basically physical beings, and that the mode of understanding the physical is basic with respect to the mode of understanding the mental, the notion of rationality cannot be immediately applied in explanations of an individual's behaviour. The difficulty which faces the physicalist is that whereas rationality is a constitutive feature of the mode of explanation

that enables us to understand an individual as acting in light of what he thinks, it is not a constitutive feature of the basic mode of explanation that enables us to understand an individual's behaviour as having been caused by an internal neurophysiological state or event. Or in other words, the difficulty which faces the physicalist is to explain how it can be possible to attribute thoughts to physical beings in explaining their behaviour. The attribution of thoughts to individuals depends on whether their behaviour can be described in terms drawn from the system of concepts that is governed by the constitutive ideal of rationality, and the physicalist wants to face up to this difficulty by arguing that the attribution of thoughts to individuals depends on the possibility of *redescribing* or *interpreting* their behaviour in terms of a system of concepts, which can only be assumed to be in place once a prior context of human interaction has been presupposed.

Davidson (1984a), for instance, argues that the context of human interaction must be presupposed as a necessary precondition for the attribution of thoughts to individuals because it is only within this context that it seems to be possible to redescribe their behaviour in the appropriate terms. His argument can be rehearsed in two main steps. The first step in Davidson's argument is to defend the claim that the notion of rationality can only be in place in explanations of an individual's behaviour where we can say that the individual has a grasp of the distinction between objective truth and falsehood. The central idea is that having a thought requires there to be a background of beliefs which identify the thought by locating it in a logical and epistemic space. Having the thought that the candle has blown out in the next room requires having the beliefs that there is a candle in the next room, that the candle in the next room was lit, that a candle is an enduring physical object which does not go out of existence when it is no longer being perceived, that it has a flame which can be extinguished in a draught, and so on and so forth. The important point to realise is that Davidson considers having a belief to require appreciating the contrast between true belief and false, such that an individual who can be said to have the background of beliefs, which are necessary for his thoughts to have their content, is an individual who can be said to have an awareness of the distinction between objective truth and falsehood.

The second step in Davidson's argument is meant to show that an individual can only be said to have an awareness of the contrast between objective truth and falsehood if he can be said to be an interpreter of the speech of another. The idea is that belief necessarily emerges within the context of interpersonal communication, or within the context of the interpretation

of speech, because that context alone can provide the tools for making the required distinction between how things are in the world independently of how the individual takes things to be. In Davidson's view, an interpreter understands the speech of another by assigning his own sentences to the speaker as an interpretation of his utterances, and to the extent that communication succeeds, the interpreter has provided an interpretation of the speaker's utterances by providing the truth conditions of his sentences. If the interpreter knows that Kurt utters the words 'es regnet' whenever it is raining, he can use his own sentence 'it is raining', to provide an interpretation of Kurt's utterance by providing the truth conditions of his sentence.

Interpretation is therefore a matter of providing a disquotational translation of the speaker's utterance into the interpreter's own language, effectively giving the meaning of his sentence by giving its truth conditions in the form of a T-sentence: "'es regnet' is true if and only if it is raining". It is crucial to realise that the interpreter can only be successful in his task if he charitably assumes that Kurt holds sentences to be true whenever they are true, for it is on the basis of this assumption that the interpreter can be said to know that, in uttering the words 'es regnet', Kurt is uttering the sentence which can be said to express the belief that it is raining; and since what a speaker means by his utterances is held to be partly determined by what he believes, the interpreter can therefore be said to be in the position to redescribe Kurt's uttering 'es regnet' as an intentional act of saying 'it is raining' (1984b: 125-131).

What supports the claim that thought necessarily presupposes an interpersonal system of communication is the point that interpersonal communication alone seems to be capable of creating the logical space within which the notion of rationality emerges. But there are certain difficulties with the type of human interaction which Davidson considers to be appropriate in this regard, which are brought out quite clearly in his account of understanding others as a matter of interpreting their speech. Certainly, it is correct to point to the context of human interaction as a necessary precondition for thought; but it is problematic to hold that understanding what an individual thinks or means boils down to being able to provide an interpretation of his utterances by giving the truth conditions of his sentences. To begin with, it is rather strained to say that understanding what an individual means when he utters the words, 'it is raining', when in it is stormy outside and there are heavy raindrops splashing into the puddles, is an act of disquotational translation or interpretation. One might question the point of his uttering these words in so obvious a fashion, but this need not amount to the idea

that understanding what others mean is in every case a matter of providing the truth conditions of their sentences. For as Glock points out, understanding what other individuals mean involves a wider variety of reactions and responses than this account seems to acknowledge:

> Understanding utterances in one's own language is not a matter of disquotation. For being able to offer such disquotations is neither a necessary nor sufficient condition for understanding an utterance...A child may understand an utterance without having mastered the apparatus of disquotation, i.e., without being able to say 'By "Shut the door" she means shut the door'. Equally, a person who can utter such sentences cannot on those grounds alone be said to understand the utterance. Instead, understanding is manifested in shutting the door, or perhaps by refusing to do so on the ground that it is too hot, etc. What is required is the ability to react appropriately to the utterance, to grasp its implications, and to explain its meaning if challenged. (1993: 203).

It seems to me that a suitable expansion of this objection would allude to the contexts of direct and indirect human relationships as providing the setting in which the appropriateness of the response is determined. Suppose that the individual who utters the words, 'it is raining', is a farmer who has been praying for the rainy season to start early after the devastating heat of the summer. In uttering these words the farmer is expressing his relief that his crops are not going to be completely ruined by the drought; understanding what he means in uttering these words and what he is thinking about when he smiles with relief is not about knowing the truth conditions of his sentence, but about knowing what the early start to the wet season means for his capacity to supply wheat and grain to the surrounding villages. Suppose that the individual who utters the words, 'it is raining', is the leader of a mountaineering team who have just started their climb. In uttering these words the leader is expressing his concern for the climb that they have just embarked on and for the safety of his team members; understanding what he means in uttering these words and what he is thinking about when he pauses with a worried expression is again not about knowing the truth conditions of his sentence, but about knowing what the impending rain fall means for the progress of the climb. Simply by considering the differences in what is involved in understanding the utterance 'it is raining' in these two cases, it makes it difficult to retain the idea that understanding others is a matter of disquotational translation. More than this, it indicates that it is extremely problematic to diminish the importance of the wider variety of non-interpretative responses to the individual, and the contexts in which these responses are made, for such responses are also constitutive of our understanding the individual, and the wider context itself determines what counts as an *appropriate* response when he utters the words 'it is raining'.

4.2.The Underlying Account Of Human Relationships

Underlying this account of thought there appears to be a problematic account of human relationships, which might be usefully discussed at this point. There is an implicit assumption, which is driven by the physicalist's metaphysics, that every human relationship is at root a relationship between two physical beings. By this account, individuals understand each other in a personal sense if they can successfully redescribe the other's behaviour in terms of a system of concepts which are not immediately applicable to that behaviour from the outset, but which can nonetheless be applied to it on a secondary level as an achievement of interpretation. An example of the type of human interaction presupposed by this position is provided by Davidson in the following passage:

> A theory of interpretation, like a theory of action, allows us to redescribe certain events in a revealing way...a theory of action can answer the question of what an agent is doing when he has raised his arm by redescribing the act as one of trying to catch his friend's attention (1984a: 161).

An explicit example of this type of human interaction is given by Jackson and Pettit:

> Suppose I want to predict how someone's body will move on some specified occasion or under some specified conditions: where do I start?...The obvious answer is: certain observed facts about what is sometimes called raw behaviour, the physical movements our bodies make described as such, rather than, for instance, the movements described in terms of the language of intentionally characterized action. For it is the raw behaviour which we more immediately perceive through the way that it impinges on our sense-organs. (1993: 261-2).

To highlight the connection between the physicalist's causal account of the mental with his interpretationist account of human relationships, here is Child:

> Interpretationists allow that there are connections between a person's having the beliefs and desires she does and the internal causal organization she does. (What attitudes a subject has is a matter of how she can be interpreted, which is answerable to what she says and does; and at some level, we think, her saying and doing what she does results from her being causally organized in the way she is.) (1994: 9).

The interpretationist seems to assume that every human relationship is *underwritten* by a specific type of impersonal relationship, one which is entered into for the purpose of predicting and explaining an individual's behaviour in terms of a system of concepts which are not immediately applicable within the context of the relationship as it stands.[6] Suppose,

---

[6] Cockburn (1990: 80-106) suggests that a significant range of materialist views can be formulated in terms of the assumption that an interest in prediction and control is, in some sense, what is basic in our relations with each other. For if the basic way in which other people figure in our thinking is such as to facilitate prediction and control of their behaviour, then our interest in others can only be accurately

for instance, that we wanted to explain why an individual climbed up into the tree in the garden. The standard interpretationist account might run something like this. The individual climbed up into the tree because he believed the cat ran up it and he wanted to get it down again. But since we are presented with raw behavioural data whose most basic description is in physicalistic terms, our explanation which alludes to the individual's beliefs and wants must have involved redescribing that behaviour in higher level mentalistic terms. In other words, it is only as the result of interpreting that we are in the position to give such a redescription of the individual's behaviour and its causes. The task we originally began with, to causally explain the individual's behaviour when he is moving up into the tree, remains the same throughout this whole complex process. What differs is only the manner in which we are able to pick out the individual's behaviour and its causes as the result of interpreting it. Rather than being restricted to talking in terms of neurophysiological states and events, and physicalistically characterised bodily movements, we can now talk in terms of the individual's beliefs and wants concerning the cat, and his actions as they were caused by the latter.

Thus, the submerged strand in the interpretationist account seems to be that all human relationships retain the basic features of non-everyday relationships, in that they are entered into for the purpose of predicting and causally explaining an individual's behaviour, whilst the internal states and events which are picked out for this purpose are described in terms of concepts which can only come into view within the context of everyday relationships. Which is to say that every human relationship is conceived to be a non-everyday relationship which has evolved into an everyday relationship through a complicated process of interpretation or redescription.

The central problem with this is that if we assume that every relationship between individuals begins on this type of impersonal level, there seems to be no obvious way in which the demands of personal relationships could ever have become an issue for these individuals, whose basic way of relating seems to be always in terms of the demands of impersonal

---

expressed in terms of features of the non-human world of the physicalistic ontology. And much the same point is made by Hacker (1996: 425): "The idea that our ordinary psychological vocabulary is part of a rudimentary causal theory of human behaviour induces the idea that its primary function must be the prediction and control of human behaviour." This would explain a great deal. It would encourage the idea that mentalistic explanations are capable of meeting our explanatory requirements only in so far as mental events, states and processes are in fact identical with neurophysiological events, states and processes, since it is only in these terms that such mentalistic-physicalistic predictions of behaviour can be made.

relationships. To suppose that the basic form of human relationships is impersonal in the manner implied by interpretationism is to suppose that the basic motivation for entering into human relationships is to predict what others will do next and how they are likely to react in such and such circumstances. But the ability to make such predictions presupposes an understanding of human motives and intentions, purpose and reasons, weaknesses and failings of character, the way attitudes vary from person to person, and so on, which requires that we can already relate to others on an everyday level. Interpretationism seems to reverse the priorities here to bad effect. For if every human relationship starts off on a non-everyday level, we are left without an explanation of how the process of interpretation could purport to bring such notions into view. And unless these notions were already in view, it seems to me that there would be no obvious explanation of why individuals ever engaged in precisely these acts of interpretation in the first place.

The immediate consequence of accepting the interpretationist approach to human relationships is that it prevents us from appreciating the deep extent of the autonomous nature of mentalistic explanation. This is due to the fact that it prevents us from seeing that the various concepts used in expressing and directing our interests and concerns in our relationships with each other fulfil the specific roles demanded of them without implicating the explanatory resources of the physical sciences. In the interpretationist's account, mentalistic explanations cannot be credited with this deeper degree of autonomy because mental concepts are used within the contexts of what are, at root, non-everyday relationships. Within the context of such hybrid relationships, mental concepts are certainly constrained by the principles of rationality; but part of their use is to predict and causally explain an individual's behaviour, which they do by apparently redescribing neurophysiological states and events in terms of mental properties. Such uses of mental concepts therefore straddle both the mental and the physical explanatory domains, and hence incorporate features of these apparently incommensurable domains by having both rationalistic and physicalistic implications built into them.

To my mind, this fails to acknowledge the deep extent of the autonomy of the mental domain, since it fails to acknowledge the possibility that mental concepts can fulfil the tasks required of them without having to straddle both the mental and the physical explanatory domains. It seems to me that we can only acknowledge the autonomy of the mental in the deeper sense through coming to appreciate the role of mental concepts within the context of

everyday human relationships, which they can fulfil without implicating the explanatory resources of the physical sciences to complete their work. Here is how: mental concepts are used within the contexts of our relationships in expressing and shaping the everyday interests and concerns we have about each other, but since they are not put to work in the course of moving from physicalistically conceived non-everyday relationships to mentalistically conceived everyday relationships by way of interpretation, they can serve our interests and concerns without back-up from the explanatory resources of the physical sciences.

Certainly, there are aspects of human interaction that fall under the domain of the physical sciences, such as movements of limbs, contractions of the muscles and the like. But in so far as our interests are tied to our involvement in everyday relationships with other people, which constitute our most immediate and basic starting point in relating to others, the latter aspects are irrelevant. The explanatory practices constitutive of the physical sciences enjoy their own autonomy, as do the explanatory practices integral to our successful involvement in our everyday relationships. But what we are interested in within the latter context is not supported by drawing on explanatory resources which serve the interests particular to the physical sciences. To suppose that our mental concepts must pick out the same subject matter as physical concepts pick out is to suppose that our mental concepts are by themselves not capable of meeting the demands we make of them. And this in turn is connected to the supposition that mental concepts emerge within the contexts of relationships premissed on the need to causally explain each other's bodily movements by reference to reasons we have for acting in certain ways. But what interests us when the carpenter reaches for his sandpaper is not that he is correctly describable as being in a mental-physical state that causes his bodily movements, but that he has realised that the chair is not yet up to the standard expected. Our interest in this case is exhaustible within the context of his involvement in this relationship, and not with predicting his bodily movements by reference to internal neurophysiological states described in mental terms.

Chapter 5: Rationalising Explanations

1.Introduction

What I want to do in the balance of this chapter is argue that rationalising explanations are non-causal explanations of action, and I want to use this point to defend my larger claim that mentalistic explanations are autonomous with respect to physicalistic explanations in the deeper sense. I will argue, contrary to the physicalist, that rationalising explanations succeed in carrying out their work without having to rely on the explanatory resources of the physical sciences, as they would if the point of rationalising an individual's actions were to attribute the mental states that rationalise his behaviour by causing it, with the relevant causal processes occurring at the micro-physical level. The causal explanatory work of mental states, on this interpretation of rationalising explanations, necessarily implicates the explanatory resources of the physical sciences. But in my view, on the other hand, rationalising explanations need not be given this causal interpretation, since they are perfectly capable of fulfilling our explanatory requirements themselves, without depending on support from the underlying causal processes in the physical domain. Again, this is not to be taken as a denial that there are such underlying causal processes, and that these processes are explanatory of the individual's bodily movements; it is rather a denial of the claim that our mentalistic explanatory resources cannot meet the requirements we have of them unless they can be said to describe the neurophysiological states and processes operative in the physical domain. Denying this claim, and explaining how it is unnecessary, is what is required, I believe, to appreciate the extent of the claim that the mental is autonomous with respect to the physical.

2.What rationalising explanations explain

Rationalising explanations are intrinsic to our understanding of individuals as rational agents. They enable us to understand why an individual acted as he did, by citing the factors in light of which his actions can be seen to have been reasonable for him to perform in certain circumstances. We rationalise an individual's behaviour when we explain it as having been performed for some particular reason, or in spite of some other reasons that he may have had at the time. An individual cannot be said to have acted for a reason if he is unaware of it as being the reason for which he acted, nor can he be said to have acted in spite of other reasons

if he did not regard them as being reasons for not acting as he did. Even if his actions could have been explained in terms of certain reasons, unless the individual is prepared to acknowledge that he acted as he did for these particular reasons, the reasons cited in such rationalising explanations cannot be the reasons for which the individual acted. For unless the individual is aware of the reasons for which he acted, it might happen that the reason that is cited in the explanation of his action is one which the individual could not have rationally had, given the rest of what he thinks and believes.

Rationalising an individual's actions can sometimes be a complicated task, since it might involve taking into account a wide variety of factors which are not immediately obvious from the way the individual is behaving. What the individual thinks or believes is not always discernible from studying his behaviour alone, and the reason for which the individual is acting is often complicated by the fact that the individual is immersed in a variety of relationships that might well have some important bearing on how his actions are to be understood in any particular case. Rationalising explanations explain why an individual is acting as he does, but our ability to understand some individuals might therefore be frustrated by the fact that aspects of their personal psychology are deliberately hidden from us, or just plain difficult to get at, more so when our acquaintance with them is limited to only particular aspects of their lives. Where the important factors stem from aspects of the individual's life in which we may not be directly involved, it is often difficult to know straight off why the individual is acting in a certain way. Failure to appreciate the violence suffered by an individual in a particularly torrid marriage might leave it quite unclear to her work colleagues why she was so apprehensive about socialising after work. But to closer friends and relatives, who are directly involved in the relevant aspects of her life, the reasons for her apprehension are immediately obvious. Knowing her as a close friend or a relative puts certain individuals in a position to understand why she is acting as she is in certain situations, whereas these reasons might be completely missed by those individuals who know her only in a working capacity. At the same time, those who know her in a working capacity might be in a better position than some of her close friends and relatives to understand why she was so averse to the company's proposed merger with a large local business.

The point I am trying to get at here is simply that their involvement in different types of relationships with this individual puts different people in a better or worse position to understand different aspects of her personal psychology. Certain individuals are in a better

position to understand why she refuses to socialise after her work through being involved in family or friendship relations with her, whereas other individuals are in a better position to understand her business ideas through being involved in working relationships with her. So what I want to suggest is that understanding why this individual is acting as she does is a matter of seeing her action as part of her involvement in these relationships, and the rationalising explanations we offer might then be understood as expressing the interests we have in this aspect of her actions.

To be precise, my position is that understanding an individual's reason for acting as he does is understanding his actions as being rational responses to features of the situations in which he finds himself. In a large number of cases, though not in all, this means that understanding an individual's actions as having been performed for particular reasons is understanding his actions as being partly constitutive of his involvement in relationships with other people.[1] This is manifest in the fact that the identity of an individual's actions are often inseparable from the identity of the actions of those individuals with whom he is involved in these relationships, and also in the fact that the identity of an individual's actions are often inseparable from the identity of his own previous or future actions, which in turn are bound up with the identity of the previous and future actions of others. The act of giving and receiving gifts might illustrate this point. An individual's act of giving a gift to his loved one is intrinsically tied to his loved one's act of receiving the gift. It is understood as the act of giving a gift by the individual receiving the gift because they are both already involved in the types of relationships in which the act of giving a gift is sometimes warranted, and in which the act of receiving the gift is the appropriate response to make. But through their involvement in *this* relationship, the individual receiving the gift is in the position to understand this particular act of gift giving as a simple expression of love, as a plea for forgiveness for a wrong doing, or as a peace offering for a recent lover's quarrel.

It would be rather artificial to suppose that the identity of the act of giving a gift could be sharply separated from the identity of the act of receiving a gift, as if it were a free standing piece of behaviour whose significance could only be grasped through a process of interpretation. The individual does not have to interpret the behaviour of her loved one in

[1] I say 'in a large number of cases' because, as in the previous chapter, I do not want to say that solitary individuals cannot act for reasons, and I want to include actions which individuals perform on their own, such as bathing, playing solitaire, listening to music on head-phones, or going for a walk *in order to be alone*. But again, the capacity to be alone, to do things by oneself, is a determination of the capacity to be with others, and to be involved in various types of relationships.

order to understand it as an act of giving a gift, before she can be said to be in a position to respond to his act in the appropriate manner by receiving it. It might be that the individual has to work out whether the act is a simple expression of love, or whether it is an act of guilt for a particular transgression; but that is only likely to be the case where the past history of the relationship suggests reasons for doubting the sincerity of the offering. What might have to be worked out is the significance of the gift, but not that the individual's act is an act of giving a gift. That is already taken for granted as partly constitutive of the particular relationship in which the individuals are involved, which is manifest in the fact that she responds to it by graciously receiving it, politely rejecting it, or suspiciously questioning the motives behind it; and depending on which response she makes, the immediate course of the relationship will be affected in some way.

In these cases, the individual can be said to have a reason for acting in a certain way if the demands of his relationships have a propensity to solicit certain types of response from him, and the individual actually acts on a particular reason if the demands of his relationship solicit that particular response from him which he would explain by citing one or other of these demands. To illustrate this point, let me modify the example: that the passenger on the train is in possession of an invalid ticket is a reason for the conductor to escort the passenger from the train at the next station. The situation in which the conductor finds himself presents him with a reason for taking such steps, given his awareness of the obligation of passengers to be in possession of a valid ticket, and his awareness of his duty as a conductor to reinforce this particular obligation. The conductor's awareness of these demands can only be acquired through being involved in the appropriate type of social relationships, in which the identity of the act of buying a ticket is intrinsically tied to the identity of the act of granting admission. With these interrelated acts comes a variety of demands, to which both individuals must be responsive if they are to be said to have the capacity to be involved in these particular relationships. So it can be said that the conductor escorts the passenger from the train for the reason that he is in possession of an invalid ticket if, in explaining why he acted as he did, the conductor would truthfully cite the fact that the passenger is in possession of an invalid ticket as demanding that these measures be taken.

Explaining an individual's actions as having been performed for a particular reason is therefore a matter of connecting his actions to certain features of the situation in which he finds himself. The connection is forged in terms of the rationalistic and normative relations

which are in place for the individual in virtue of his conception of the situation, and the causal relations which are explanatory of his bodily movements are *extrinsic* to this being the case. In explaining why an individual acted as he did we are interested in getting at his reasons, and we do this by understanding which features of the situation strike him as particularly salient. (We are not interested in what caused his body to move in any particular way, and even if we did have this information it would not tell us what we want to know.) The fact that the passenger is in possession of an invalid ticket strikes the conductor as particularly salient given his conception of the situation, and it is in terms of his conception of the situation that his actions are connected to that fact by means of the rationalistic and normative relations which are intrinsic to the rationalising form of explanation.

This takes me to a crucial point, which raises a problem for the causal account of rationalising explanations: if it is correct to say that the reason for the conductor's action is the fact that the passenger is in possession of an invalid ticket, then it cannot also be correct to say that his reason for acting is, as the causal account of rationalising explanations has it, the belief-desire combination which causes him to act. The causal account of rationalising explanations is committed to the idea that an explanation of an individual's actions in terms of his reasons is a species of causal explanation, for in order to have a causal account of rationalising explanations, the reason for acting must be construed as that which rationalises the individual's action by causing it. Or as Davidson puts it, the reason must be a rational cause of the action; it is the belief and desire in light of which the action can be seen to be reasonable (1980b: 233).

What I am suggesting is that the reason for acting is not the belief and desire which the individual has at the time of action, and which work together in some mysterious way to cause the behaviour that the reason is supposed to explain; rather is the reason for acting given by *what* the individual thinks or believes and *what* the individual desires on that occasion. Here is what I mean. That the bulls are too lean for selling at the market is the farmer's reason for fattening them up. Certainly, this fact would not present the farmer with a reason for doing any such thing if he did not have the relevant beliefs and desires. But what the reference to the farmer's beliefs and desires does is explain why this fact is a reason for *him*. They themselves do not constitute his reason. Thus, I feel that the causal account of rationalising explanations involves a basic slippage, to which the physicalist has been attracted because it allows him to retain a physico-causal account of behaviour, whilst at the

same time it appears to let him retain the notion of rationality as integral to that account, thus giving him the possibility of arguing that the mental is nonetheless autonomous with respect to the physical. But what I am claiming, on the other hand, is that rationalising explanations do not incorporate the notion of causality as a necessary component. However, this is challenged by certain arguments which are thought to motivate the claim that the concept of acting on a reason *must* contain within itself two ideas: the idea of rationality and the idea of a cause.

3.Why must rationalising explanations be causal explanations?

3.1.Accounting For The Force Of 'Because'

The causal account of rationalising explanations has often been appealed to as the only means of explaining the relation that obtains between the reason and the action. It is sometimes motivated on the strength of the point that merely averting to the fact that an individual has a reason for acting is not a sufficient condition for his actually acting on that reason. Davidson's classic argument for the causal account is this: although rationalising explanations justify an individual's actions in the sense that they cite the beliefs and desires in light of which the individual's actions can be seen to be reasonable, they must also be causal explanations in the sense that the beliefs and desires which rationalise the individual's actions must do so in virtue of being their causes. Giving a reason explains an action in the sense that it allows us to fit it into a wider context, but this still leaves us without an account of how the reason explains the action, and it fails to touch on the question of whether the reason which is cited in explanation of the action is in fact the reason why the individual acted. So the explanatory efficacy of reasons can only be accounted for on the assumption that reasons are causes, otherwise rationalising explanations will fail to account for the force of the 'because' which is operative in these cases (and what better way to do this than through the slippage outlined above?). The problem can then be put in the following way: it is one thing to cite the beliefs and desires which reveal an individual's actions to be reasonable for him to perform, but it is another thing to show that having these beliefs and desires actually explain the individual's actions. And as Davidson concludes:

> A desire and a belief of the right sort may explain an action, but not necessarily. A man might have good reasons for killing his father, and he might do it, and yet the reasons not be his reasons in doing it...so when we offer the fact of the desire and belief in explanation, we imply not only that the agent

had the desire and belief, but that they were efficacious in producing the action. Here we must say...that causality is involved (1980b: 232).

As I have already pointed out, this argument is less than convincing. It surely cannot be on these grounds alone that "we must say" that causality is involved. What motivates the causal account of rationalising explanations, at least as far as this argument is concerned, is that there seems to be no other way of explaining the relation between the reason and the action when the reason in fact explains the action. Certainly, it is correct to say that the man might have a good reason for killing his father and yet not kill him for that reason. But this cannot be sufficient grounds for concluding that the reason on which the individual acts must be the one which causes it. If it turned out that the man shot his father accidentally at archery practice, when the fact that his father had a vast estate gave the man a good reason for killing him anyway, then the man did not act on the latter reason because giving that reason in explaining his actions would be incorrect.

The point is straight-forwardly grammatical: the reason on which the individual acts is the reason which gives the correct explanation of his action, and the correct explanation is the one which the individual would truthfully give if asked. The individual's own truthful explanation of his actions is the criterion by which we determine the reason for which the individual acted as he did. The correct explanation is that the man fired his arrow toward the target, and the man's father happened to walk into its path just as the arrow was fired. So although the man killed his father, and although he certainly had a reason for killing him, he did not in fact kill him for that reason. This would give an incorrect explanation of the man's actions, and hence it would fail to give the reason for which the man acted as he did. But to insist that the latter reason would give an incorrect explanation of the individual's actions because it did not cause them would be to assume from the outset that the causal account of rationalising explanations were true. In order to establish this conclusion, which is by no means obvious from these considerations, further argument is required.

3.2.The Completeness Of Physics

There is a deeper argument underlying the causal account of rationalising explanations, which is more firmly rooted in the physicalist's explanatory framework. The causal account of rationalising explanations is motivated on the strength of the completeness of the physical sciences, which is the principle that there is a complete and deterministic explanation of

everything that exists, and hence of everything that happens, in terms of the explanatory resources of the physical sciences. There is no need to go beyond the physical domain to explain some physical occurrence, such as the bodily movements of the individual who is acting for some reason or other. The occurrence of the individual's bodily movements are completely and exhaustively explained in terms of the occurrence of antecedent physical events; so if mental states or events are to stand a chance of being explanatory of an individual's behaviour, then they must somehow merge with these physical states and events, either through being realised by them, or through being mereologically composed out of them. But this means that once reasons are construed as 'belief-desire pairs', rationalising explanations must be supported by underlying physicalistic explanations, since the mental states and events cited in these explanations are not in the position to make any difference to the individual's behaviour unless they are in fact physical states and events.

This is enough to motivate the causal interpretation itself, but let me point out an immediate consequence of this argument that will help to reinforce the conclusion. For any event, there must be an explanation of its occurrence that cites its physical properties, locating that event within the network of causal relations. Certain events can also be explained in rationalistic terms, by citing their mental properties, but there cannot be a rationalistic explanation of an event for which there is no physical explanation. For any two events which differ with respect to their rationalistic explanations, there must be a difference with respect to their physical explanations, which will thereby register a difference in the causal relations that these events can enter into. It follows from this that if two events differ with respect to their rationalistic explanation, they must also differ with respect to their physicalistic explanation, which is to say that for any rationalistic explanation to be true of an event, there must be some physicalistic explanation on which this rationalistic explanation depends. From this, the desired conclusion seems to follow without too much difficulty: rationalistic explanations must be a subspecies of causal explanation.

In practical terms, what this means is that if the explanation of the man's killing his father is that he believed his father had a secret pot of money and he wanted to inherit that pot of money sooner rather than later, then there must also be some causal explanation of the man's actions on which the truth of this particular rationalising explanation depends. If the man did not act for this reason, but accidentally killed his father anyway, then there must of necessity be a different causal explanation of the man's actions. So at a basic level, what distinguishes

an action which is performed for one reason, from an action which is performed for another, must be the truth of the causal explanation of the occurrence of the individual's behaviour, on which the rationalising explanation depends.

Despite the fact that this argument looks to be much more sophisticated, it is doubtful whether it is any more successful than the previous argument. It is, however, much more difficult to criticise; but a useful starting point might be to question whether the causal account of rationalising explanations, as it is motivated on the strength of this argument, is internally coherent. If it is correct to say that the notions of rationality and causality must be built into the concept of acting on a reason, then it seems that it must also be correct to say that the autonomous nature of mental properties does not prevent them from being causally relevant. But from within the non-reductivist's metaphysical framework, it can be argued that autonomous mental properties are in danger of being causally irrelevant.

This gives rise to the problem of mental property epiphenomenalism, which is the problem that the instantiation of mental properties in events is irrelevant to the causal relations that these events can enter into. Which causal relations events can enter into is determined exclusively by their physical properties, leaving no causal work for their mental properties to carry out. This is to say that had an event failed to instantiate any of its mental properties, once its physical properties had been fixed, no difference would have been made to the causal relations that that event could have entered into. It could have entered into the same causal relations, regardless of whether it had instantiated any of its mental properties or not. On a closer inspection, this problem can be seen to divide into two: first, once an event's physical properties have been fixed, there is no causal work that its mental properties can do; second, once an event's physical properties have been fixed, there is no causal work that its mental properties need do.

4.The problem of mental property epiphenomenalism

4.1a.The Causal Irrelevance Of Non-Nomological Properties

Honderich (1982: 60-62) raises a problem of the first type. Suppose that the event of putting pears on a scale causes the pointer to register two pounds, and suppose that the events are described in exactly these terms. It is not necessary that every property of these events will be

picked out by just this description, and some of the properties which are not picked out by this description may be the properties which are in fact relevant to the causal relation that holds between these events. As Honderich points out, it is only in virtue of certain of their properties rather than others that events can be said to be the causes they are. The events cannot be said to be in lawlike connection in virtue of the first event being the event of putting something green and French on the scale, and in virtue of the second event being the event of the pointer's registering two pounds. As far as determining the causal relation between these events is concerned, the latter properties are irrelevant. But although the events are not brought under a law in virtue of being described in exactly these terms, it does not mean that the event of putting pears on the scale did not cause the pointer to move. What it means is simply that the failure of certain properties to figure in lawful generalisations entails that it cannot be in virtue of these properties that the events enter into the causal relations they do.

This creates the following difficulty for the causal theory of rationalising explanations: given that events which instantiate mental properties also instantiate physical properties, and that the causal relations which hold between events do so in virtue of the fact that they instantiate law governed properties, how can it be the case that the mental properties of these events are relevant to the causal relations that they enter into? Suppose that we explain an individual's behaviour of walking to the water fountain by saying that he wanted a drink, and that he believed that he would be able to quench his thirst at the fountain. What this explanation seems to imply is that the event which causes the individual's behaviour instantiates the mental properties cited in the rationalisation of his action. But since the event cannot be said to cause the individual's behaviour in virtue of instantiating any of its mental properties, it follows that the mental properties cited in this explanation are irrelevant to the fact that the individual walked to the water fountain. So there now seems to be no non-arbitrary explanation of the fact that the event which caused the individual's behaviour is the event which instantiates the mental properties cited in the rationalisation of his action. Or in other words, there now seems to be no non-arbitrary explanation of how the individual's reason for walking to the fountain caused his walking to the fountain; yet it was supposed to be the merit of the causal account that it alone could provide an account of the relation between reasons and actions.

4.1b.Causal Explanatory Exclusion

Kim (1993c: 351-355) raises a problem of the second type. He argues that the causal account of rationalising explanations, as it is set out within the framework of non-reductive physicalism, flounders over the causal explanatory exclusion problem. This is basically the problem that the causal role of an event's mental properties is in danger of pre-emption by the causal role of its physical properties. Or in other words, once an event's physical properties have been fixed, there is no causal work that its mental properties need do. Suppose that mental property M is causally efficacious with respect to mental property M*, and that on a given occasion, the instantiation of M causes the instantiation of M*. But given that mental property M* is realised by physical property P*, there are now two different answers to the question why mental property M* is instantiated: on the one hand, the answer could be that M* is instantiated because it was caused by the instantiation of M; on the other hand, the answer could be that M* is instantiated because it is realised by the physical property P*, which happens to be instantiated on that occasion. So if the instantiation of mental property M, is to have caused the instantiation of mental property M*, it could only have done so if it had caused physical property P* to be instantiated, since the instantiation of P* is already nomologically sufficient for M* to be instantiated. But this presupposes the possibility of downward causation, from higher-level mental properties to their lower-level physical realisation bases.

Now suppose that on a given occasion the instantiation of M causes the instantiation of P*, as we can grant if we allow the possibility of downward causation. But given that mental property M is realised by physical property P, and that the instantiation of P therefore explains why M is instantiated, it follows that the instantiation of P itself is sufficient for the instantiation of P*. The question which arises now is why should we not take P as the cause of P*, and disregard M as an epiphenomenon? If the instantiation of P* has a complete and sufficient physical cause in the form of the instantiation of physical property P, as the principle of the causal closure of the physical domain demands, there is no reason for taking M to be the cause of M*, since M is realised by P, M* is realised by P*, and P completely explains P*. All the causal work is carried out at the level of physical properties, leaving no causal work for mental properties to do. But Kim insists that the causal efficacy of mental properties might still be secured, if the principle of 'causal inheritance' is adopted. By this principle, if mental property M is instantiated on a given occasion by being realised by

physical property P, then the causal powers of M are identical to the causal powers of P; all causal relations are therefore implemented at the microphysical level, such that the causal powers we attribute to mental properties are inherited from the causal powers of physical properties.

Again, in practical terms, what this amounts to is that the mental properties cited in the explanation of the individual walking to the fountain have no causal work left to do once the physical properties of the behaviour-causing events have been fixed. There is no causal work left for mental properties to do, unless they are permitted a downward causal role with respect to the physical properties of the individual's behaviour. But since there is already a complete explanation of the individual's behaviour in terms of the instantiation of the physical properties of the behaviour-causing event, granting autonomous causal powers to mental properties is not warranted, since it simply results in the causal over-determination of the individual's behaviour. This has an unhappy consequence for the causal theory of rationalising explanations: if mental properties fail to have their own causal powers, there is little reason to consider mental properties to characterise an autonomous domain; and if mental properties cannot be credited with the autonomy that the non-reductivist demands, this will put pressure on the idea that the rationalising mode of explanation can be said to be a species of causal explanation, which is nonetheless autonomous with respect to physicalistic explanations.

This is unacceptable for the causal account of rationalising explanations, in so far as one of its central aims is to find a way of reconciling the autonomy of the mental with the fact that the physical sciences provide a complete explanation of the individual's behaviour in terms drawn exclusively from its own field. Once the completeness of the physical sciences is assumed, there is no room left for rationalising explanations to get a grip as an autonomous form of causal explanation. All the causal work is carried out at the lower-level of physical properties, which means that mental properties can only be said to be causally efficacious if we allow their efficacy to be inherited or derived from that of physical properties. But this leaves the higher-level mental properties with no autonomous causal role to play. The causal account of rationalising explanations appeared to show how both modes of explanation could co-exist when they take as their subject matter the causation of an individual's behaviour; but the two objections just considered cast doubt on the possibility of keeping them both in place

together, and indeed, on the very need to retain the rationalising mode of explanation once the completeness of the physicalistic mode is acknowledged.

4.2a. The Causal Efficacy of Higher-Level Patterns

That mental properties are relegated to the status of epiphenomena within the framework of non-reductive physicalism seems to reveal a deep incoherence in that position. Part of the merit of non-reductive physicalism is that it seemed to secure the autonomy of mental properties, despite the fact that they are grounded in the physical structure of the world. But before setting non-reductive physicalism aside on the above objections, it is reasonable to consider whether they can be answered. The problem facing the non-reductive physicalist is that the causal powers of higher-level properties in general seem to be completely derived or inherited from the causal powers of the underlying physical properties, whose arrangements can be said to realise the higher-level properties on a particular occasion. For if higher-level properties are realised by particular arrangements of lower-level physical properties, then it does seem to follow that the causal powers of higher-level properties are nothing over and above the causal powers of lower-level physical properties. But it might be replied that the charge of mental property epiphenomenalism cannot be sustained, if it can be demonstrated that mental properties can in fact exert some form of causal influence over physical properties.

This is the line taken by Van Gulick (1993: 250-252). He argues that the charge of mental property epiphenomenalism is not as serious a threat as it initially seems. His reply to the charge of epiphenomenalism is this: it does not immediately follow from the fact that mental states and events are realised by composites of physical states and events, that mental states and events cannot exert some form of causal influence over the physical states and events of which they are composites. It does not follow because the causal powers of higher-level states and events are not solely determined by the causal powers of their constituent parts, but by the causal powers of their constituent parts *together* with their particular organization. The patterns that are picked out by the predicates of the special sciences in general are stable and recurrent features of the world, which are preserved as such regardless of the changes and alterations taking place among the lower level constituents of these patterns. Van Gulick's central point is that the existence of these patterns in the world determines that the particular constituents of their instances are organised and recruited as they are. This has the effect that

higher-level patterns are responsible for selectively activating only some of the causal powers attributable to their constituents parts. That is, the physical constituents of these patterns may have many causal powers, but only some subset of these causal powers will be activated in a given situation. Through being organised into this particular higher-level pattern, certain subsets will be caused to be activated and others will not, and the higher-level pattern, therefore, can be said to exert its own causal influence over its physical constituents by selectively activating certain of these subsets rather than others.

If this argument works, it suggests reasons for thinking that the higher-level patterns picked out by the predicates of the special sciences are capable of exerting some of their own causal powers, which are not entirely derived from the causal powers of their constituent parts. Van Gulick argues that the activity of a reagent can be affected by the presence of a catalysing enzyme that forms a composite with the reagent, and this apparently weakens the force of the epiphenomenalist's charge. However, even if it is correct to say that when biochemical states and events combine to form composites they acquire their own autonomous causal powers, we do not yet have reason to suppose that this can throw light on the present issue. It does not show that if the composite is a mental event (the intelligibility of which has already been put in doubt, given the problem with fusions) it can exert an autonomous causal influence over the activity of its constituent parts. It is one thing to argue that biochemical composites can have a causal effect on their constituent parts, but this does not help us understand what it means to say that mental states and events can exert some causal influence over the physical states and events which purportedly form their constituents parts. The trouble is that whereas we readily understand the principle behind the idea that higher-level biochemical composites have lower level biochemical states and events as their constituent parts, we do not so readily understand the principle behind the idea that mental states and events have constituent parts in the same sense. But this is exactly what we do no need to understand if the analogy between mental states and events and higher level biochemical composites is to help us understand how autonomous mental properties can exert underived causal powers over physical properties.

4.2b.Non-Reducible Supervenient Causation

A different reply is suggested by Enç (1995: 178-180). He argues that there is a species of supervenient properties whose causal efficacy cannot be fully accounted for by the causal role

of their microphysical base. The properties in question are locally supervenient properties that are associated with certain globally supervenient properties. It is because certain locally supervenient properties can be said to "carry" certain globally supervenient properties that their causal efficacy cannot be exhaustible by the causal role of their microphysical base. Thus, we seem to have found a role that mental proeprties need to fulfil. Here is how the argument works. For each representation of a state of affairs, there is a structure which embodies that representation, and which locally supervenes on some neurophysiological properties of the brain. Since this same structure could have been instantiated in a baby who does not yet have any beliefs, Enç holds that it cannot be identical with the property of being a representation of a state of affairs. The relation between the macro structure of the brain, and the property of being a representation, is rather one of carrying. The property of being a representation of a state of affairs includes in its supervenient base properties outside the confines of the individual's body. So according to Enç, the neurophysiological state of the brain only has the property of representing this state of affairs in virtue of the fact that it is a realisation of the structure that would be ideally caused by this state of affairs.

What this means is that although the neurophysiological state of the brain is itself sufficient to determine that a given macro structure is instantiated, it is not sufficient to determine that this structure constitutes a representation of the state of affairs in question. That this structure constitutes a representation of the state of affairs cannot be derived from the fact that it locally supervenes on some neurophysiological state of the brain. The neurophysiological state of the brain can only be said to constitute a representation in virtue of the fact that it is a realisation of some macro structure that would have been caused by the relevant state of affairs under certain specifiable ideal conditions. Now by Enç's reasoning, a piece of behaviour is intentional behaviour only if it has been caused by a representational state. The intentional behaviour of running away from a tiger, for instance, is constituted as such only if it has been caused by the representation of a tiger. The neurophysiological state of the brain will be sufficient to cause the individual's bodily movements when he is running away from the tiger, but that the piece of behaviour is, as Enç puts it, "an intentional running away", is not fully determined by the fact that it was caused by this neurophysiological state of the brain. That the piece of behaviour has the property of being intentional is dependent on whether it has been caused by the neurophysiological state that realises the relevant representation, but the neurophysiological state realises the relevant representation only in virtue of its relations to factors outside the individual's body.

Enç argues that when a globally supervenient property (being a representation of a tiger) is carried by a locally supervenient property (being a given macro state of the brain), a novel causal structure is created, which is different from the causal structure obtaining at the micro-physical level. Because the properties of being a representation of a state of affairs, and being a piece of intentional behaviour, are globally supervenient properties that are carried by certain locally supervenient properties, a significantly different level of causal relations obtains between these events. This higher level of causal relations is superimposed on the causal relations already existing at the micro level, in virtue of the fact that the properties which locally supervene on the micro properties carry certain globally supervenient properties, whose causal efficacy cannot be fully accounted for by the causal role of the lower level micro properties. By Enç's lights, the causal relations at this level are *nonreducible* to the causal relations at the micro level, and this seems to secure his claim that mental properties play a causal role over and above the causal role played by microphysical properties.

However, this seems to imply that a single event can be involved in a large number of causal relations at the one time, corresponding, more or less, to the number of descriptions that can be made of it. It seems to imply that the neural event stands in a causal relation to the contraction and expansion of the individual's muscles, and that the same event, described in terms of certain of its macro properties, stands in a causal relation to the individual's bodily movements. This might not seem problematic. But I think the problem rather lies with the next claim, that the same event again, when described in terms of its representational properties, stands in an *independent* causal relation to the individual's intentional behaviour. It is not clear to me that Enç has identified an autonomous level of causal relations that is independent of, and non-parasitic on, the causal relations at the micro level. Even if we agree, for the sake of the argument, that there are a number of different and independent causal stories to be told about the same neurophysiological events, as Enç insists, this does not seem sufficient to warrant his claim that, corresponding to a nonreducible mental description of these events, there is an *independent causal relation* that is nonreducible to certain lower level causal relations.

The independence of the higher level causal relation from the lower level causal relation is thought to have been secured by the claim that the causation of the individual's behaviour by

certain neural events is not sufficient to make it the case that the behaviour is intentional. This claim is certainly correct; but in so far mental states and events are identical to the physical state or event, this observation only seems to secure the need to recognise an autonomous level of description. It does not seem to secure the fact that there is an independent level of causal relations between these events, corresponding to their independent higher level descriptions. For although we might not be able to derive the complete set of an event's higher level properties from a description of that event in terms of its causal relation to some other event, it does not follow that we have to assume that these underivable higher level properties have been independently caused by the higher level properties of the first event. For this reason I am doubtful that this argument is any more successful than the previous argument in securing an autonomous level of mental causation, and again the prime reason for this failure seems to be that the token mental event is assumed to be identical with some lower level token physical events.

5.Securing the autonomy of mental causation by rejecting the principles of physicalism

The arguments considered against the causal account of rationalising explanations seem to imply the denial of the fact that what an individual thinks can sometimes be causally explanatory of what he does. I do not want this denial to follow from the non-causal account of rationalising explanations that I have been developing, so I need to find a way of making sense of mental causation that is consistent with my position. This should not be too difficult, since mental causation only seems to be put in doubt if it is understood in terms of the principle of the nomological character of causality and the completeness of physical theories. In terms of these principles, causation is an extensional relation between two events, which holds independently of the various explanatory frameworks in terms of which the events can be described; but the principles require that whenever this relation does in fact hold, there must be a closed and comprehensive framework whose laws govern the events under certain of their descriptions. This is what has been seen to cast doubt on the autonomy of mental causation. But there should be no obvious reason to deny that the autonomy of mental causation can be secured if it can be understood independently of these principles; nor should there be any obvious reason to deny that the autonomy of mental causation can be made consistent with the non-causal account of rationalising explanations.

Baker (1993: 92-4) suggests a way of understanding mental causation independently of the principles of physicalism, which involves reversing the priority of causation and explanation, and then grounding the notion of causation in an array of counterfactual facts constitutive of our explanatory practices. In terms of this reversal, the notion of causation comes to be viewed as an explanatory concept which figures in our explanations of everyday happenings, and causes come to be viewed as that which is cited in these explanations. More specifically, causal explanations are explanations which are given in answer to 'why?' questions, such as 'why is the traffic so snarled today?', and causes are cited in answers to these questions, such as 'road works at the interchange'. Baker's recommendation is that the notion of causation should therefore be analysed in the following way: (i) if c had not occurred, then other things being equal, e would not have occurred; (ii) given that c occurred, then other things being equal, e was inevitable. Applied to explanations of actions, Baker wants to say that Jill's thinking that she left her keys in the bookstore causes her to return to retrieve them in virtue of the following explanatory facts: if Jill hadn't thought that her keys were in the store, then other things being equal, she wouldn't have returned, and given that she did think that her keys were in the store, then other things being equal, her returning to the store was inevitable.

This reversal allows us to say that what is relevant to settling whether a causal relation holds is simply whether certain counterfactual claims are true, rather than whether the events have descriptions under which they instantiate a strict law. There is no need for causally related events to be governed by strict laws because there is no need for causal explanations to cite anything like the complete causes of the type required by the physicalist's principles. Something like this idea is worth developing, since it completely avoids the troubles with mental property epiphenomenalism which threaten to undermine the non-reductivist's account of mental causation. Whether Jill's thinking caused her to act as she did would not depend on whether non-nomological properties can be said to be causally relevant, but rather on the truth of the counterfactual claim that had she not thought what she did, she would not have acted as she did. However, it seems to me that there is a problem with Baker's position, which prevents me from taking full advantage of it. There seems to be no distinction drawn between answers to 'why?' questions which are genuinely causal explanations, and answers to 'why?' questions which are not causal explanations, such as 'you gave me a fright, that is why I jumped', and 'there was somebody standing in the shadows, that is why I didn't go any further'.

If Baker's methodological recommendation to reverse the priorities of causation and explanation is to be exploited as a way of securing the autonomy of mental causation, and its consistency with the non-causal account of rationalising explanations that I have developed, there will have to be a way of distinguishing between explanations that cite mental causes and explanations that do not. It simply cannot be acceptable for my purposes to assimilate all explanations of an individual's actions in terms of what he thinks to causal explanations, since this move overlooks the different implications for certain of our relationships, that come into view through explanations that cite mental causes, and explanations that cite reasons. Anscombe (1968) might be able to help me out here. She agrees that mental causes are cited in answers to 'why?' questions, but the important point to note is that she puts a restriction on the type of question that can be said to be seeking a mental cause in its answer. It is noteworthy that the restriction she places on what can count as a mental cause is to be understood in terms of our interests in explaining a particular aspect of an individual's actions. A mental cause is simply what would be cited in describing what went through the individual's mind that led up to and issued in his action, most commonly, but not necessarily, when the individual is in a state of excitement or agitation: "The martial music excites me, that is why I walk up and down", "What made you sign the document at last?- The thought: 'It's my duty' kept hammering away in my mind until I said to myself, 'I can do no other', and so signed." (1968: 76).

In contrast to this, what characterises explanations of an individual's actions that cite reasons is that they illuminate aspects of the individual's situation as he conceives it. What is of interest is not what happened to be going through the individual's mind that led up to his action, but whether the individual's response to his situation was warranted or appropriate. So if the request for an explanation is expressing an interest in the latter considerations, then what is given in response is the individual's reasons for acting as he did; but if the request for an explanation is expressing an interest in what was going through the individual's mind that led up to his action, then what is given in response are the mental causes of the individual's action. This is a useful way of showing how the non-causal account of rationalising explanations can be consistent with the possibility of explaining an individual's actions in terms of mental causes. I think it is important to note that both types of explanation are autonomous, in the sense that their intelligibility stems from their role within the contexts of human relationships. This is more evident with regard to a request for an explanation in terms of an individual's reasons, but it can also be said of a request for an explanation in terms of

mental causes. Understanding what caused the individual to sign the document enables us to understand that he was acting under pressure, that he did not calmly come to the decision to sign it, and that our attitude toward the individual ought to be less severe than it might have otherwise been.

The mode of explanation that cites mental causes is autonomous with respect to the mode of explanation that cites physical causes, and its autonomy is to be understood in the same way as the autonomy of the rationalising mode of explanation.[2] Mental states can be granted causal efficacy without having to be realised in physical states, since their causal efficacy is to be understood in terms of the way in which they figure in our explanations of the individual's actions. Explanations of the individual's actions which reveal them to have been performed under certain circumstances, such as those mentioned above, can be legitimately described as causal explanations. Mental causes can be said to make a difference to the individual's behaviour, or simply to his current thoughts, without implicating the explanatory resources of the physical sciences, and the differences that mental causes make are manifest in the mentalistic explanation given. That an individual's thoughts are construed as mental causes is registered by the way in which they figure in our explanations, as excusing the individual's sudden decision to act as he did, as acknowledging his diminished responsibility, as explaining how he suddenly came to be thinking about a girl he once met in Vienna when he started off thinking about how quickly the flowers had wilted in the garden, as describing what was going on in his mind that led up to and issued in his action, and so on, and so forth.

The important point here is that our concept of mental causation does not implicate the explanatory resources of the physical sciences, since the explanatory work required of this concept is successfully carried out whilst remaining within the network of explanations appropriate to the mental domain itself. It is only on the assumption that mentalistic and physicalistic causal explanations explain the very same thing, namely, the individual's bodily movements, that we need to assume that mental states and events are identical with physical states and events; and it is only if we make this assumption that we are forced to argue that the explanatory work required of the concept of mental causation cannot be carried out without drawing on the explanatory resources of the physical sciences. But if it is correct to

---

[2] As Wolgast (1998: 30-31) correctly points out, there is a difference between understanding the causal processes responsible for bringing about bodily movements, and understanding mental causes, since the latter requires our *personal* understanding and *experience* of human nature, whereas the former does not, and the mental causes characterise the individual's actions, whereas the physical processes do not.

say that the concept of mental causation meets our explanatory requirements within the network of human relationships, and if it is correct to say that our interest in talking about mental causes differs from our interest in talking about the physical causation of an individual's bodily movements, then we can say that mental causation is in this sense autonomous with respect to physical causation.

6.Conclusion

Understanding others as rational agents is a matter of understanding their actions as being responses to the reasons presented to them in different situations. Although the actions that individuals perform are not in every case actions which are partly constitutive of their involvement in some human relationship, it seems to me that their responsiveness to reasons cannot be completely separated from their responsiveness to the demands of human relationships. Understanding others as physical beings, on the other hand, is a matter of understanding their behaviour as subject to determination from neurobiological states and events. But whilst it is certainly correct to state that there are natural causal relations of some sort between the mental and the physical domains, it does not seem obvious that both modes of understanding must take as their common subject matter internal behaviour-causing states and events. This only becomes obvious when both modes of understanding are forcibly merged together in a manner appropriate to capturing the sense in which the physicalist's explanatory resources are all encompassing. But it does not follow from the fact that there are such natural relations between the mental and the physical domains that the physicalist's explanatory resources must be able to tie down the mental domain in the manner required.

The problem is that once the priority of the physical is assumed, and once the explanatory resources of the physical sciences have been extended beyond their own proper domain, there no longer seems to be any room left for an autonomous mental domain. The only solution is to think of the mental as somehow identical with, or realised by, the physical, so that mental states and events can be granted the causal powers they seem to have been denied, thereby allowing them to make some causal difference in the physical world. For unless mental states and events are conceived in this manner, the concern is that they cannot be explanatory of an individual's behaviour. So it seems that the only way to secure the effectiveness of our mentalistic explanations, once the metaphysics of physicalism have been assumed, is in fact

to compromise their full autonomy by insisting that they can meet the demands made of them only if they are underwritten by the explanatory resources of the physical sciences.

But rather than granting causal powers to the mental domain, it seems to me that once the completeness of the physical sciences has been accepted, and once mental events are identified with physical events, there seems to be no warrant to grant autonomous causal powers to the mental. The causal powers granted to the mental turn out to be completely derived from the causal powers of the physical, leaving mental properties with no causal work to do. The upshot of this is that not only has the mental domain been denied its autonomy, but that the mentalistic explanations we employ on an everyday basis seem to have been deprived of their effectiveness. It seemed at first that mentalistic explanations had to be merged with physicalistic explanations in order to ensure that mentalistic explanations could meet the demands made of them; but what has actually happened, or so I have been arguing, is that the merging of our explanatory resources in this manner has resulted in mentalistic explanations having no genuine explanatory role to play at all.

I have been arguing that the explanatory resources appropriate to understanding others as rational agents do not have to draw on the explanatory resources of the physical sciences in order to successfully meet the everyday requirements we have of them, since what they explain is the individual's responsiveness to the demands of the situations in which he finds himself, and they can do this without relying on support from the explanatory resources of the physical sciences. Certainly, the individual's capacity to act for reasons causally presupposes that there are underlying physical processes bringing about changes in his bodily movements; but the mentalistic explanations we make of the individual are not geared toward rationalising his behaviour by citing the mental states and events that causally explain it. It seems to me that mentalistic explanations do not require a built-in causal component in the way that only becomes necessary once mental states and events are identified with physical states and events. Instead, mentalistic explanations are rather geared toward rationalising an individual's behaviour by allowing us, in a large number of cases, to see his actions as partly constitutive of his involvement in certain of his relationships. But again, not all actions will be of this type, since we perform actions when we are on our own which are not obviously tied to our involvement with other people, as when we go for a walk to relax or keep fit, or when we sing a song or whistle a tune just for the sake of it; and we can readily agree that certain individuals, those seeking absolute solitude for the purpose of meditation, or those

unable to cope with the pressures of modern society, may have completely withdrawn from their involvement in relationships with other people altogether. So in certain cases, understanding others might not be a matter of seeing their actions as partly constitutive of their involvement in some relationship; but this does not affect the point that what is at issue is our *personal* understanding, that we draw on our experience of human nature, and that the mentalistic explanations thus appealed to can work successfully without implicating the explanatory resources of the physical sciences.

The pivotal point in the causal account I have been concerned with seems to be the assumption that mentalistic and physicalistic explanations take a common subject matter, that the subject matter picked out by mentalistic explanations is the very same as the subject matter that is picked out by physicalistic explanations, with the only difference being the manner in which this happens. This assumption is motivated by the acceptance of the metaphysical framework of physicalism, which forces us to find a means of grounding the mental in the physical structure of the world. The fear is that unless mental properties can be shown to be realised by physical properties, unless mental states and events can be shown to be identical with physical states and events, and so on, the mental domain will represent a break down in the completeness of the physical sciences. Everything which exists is physical, and as such everything which occurs can be given a complete and deterministic explanation by means of the explanatory resources of the physical sciences. But unless the mental is somehow integrated into this deterministic network, it would have to be admitted that the explanatory resources of the physical sciences could not explain everything which exists. It is the need to make good the principles of physicalism that motivates the conception of the mental as identical with the physical, and it is this conception that prevents us from appreciating the deeper extent of the autonomy of the mental. That extent can only be appreciated by acknowledging that the explanatory resources of the mental domain can work successfully and completely without implicating the explanatory resources of the physical domain.

Chapter 6: Thoughts Externalised

1.Introduction

Externalism is the view that what an individual thinks is subject to determination from factors that lie outside the physical boundaries of the individual's own body. Which factors are considered to be relevant seems to depend very much on which factors figure prominently in one's account of the individual's relationship to his environment and to the other people with whom he lives his life. An account of the individual's relation to his environment in terms of the brute causal relations that obtain between him and the objects and events which he experiences will tend to hold that it is the nature of the physical environment that determines the content of his thoughts. An account of the individual's relation to his environment in terms of his participation in a linguistic community will tend to hold that it is the meaning of the terms in the linguistic community that determines the content of his thoughts. Which factors are relevant to the determination of an individual's thoughts is important. It not only has a bearing on the account that one can give of the individual's understanding of his own thoughts, it also has a bearing on the account that one is able to give of the explanations employed in understanding other individuals as rational agents.

That thoughts are individuated externalistically can be used to put pressure on the assumption that mental states and events are identical with internal physical states and events of the individual's body. The externalistic individuation of thought complicates matters by pointing out that what an individual thinks can be said to be dependent on factors external to his body, which arguably makes it difficult for the physicalist to defend the identity claim. Yet it is important for the consistency of his position that he can do so, since the identity claim seems to be required within his metaphysical framework to make sense of the fact that what an individual thinks and believes is causally relevant to what he does. This presupposes that the individual's mental states and events are identical with the internal behaviour-causing states or events which fall within the field of the physical sciences. The aim of this chapter is therefore to work out an account of the dependence of thought on the individual's surroundings that will put pressure on the identity claim, and hence provide support for the argument of the previous chapter. I will begin by considering whether the versions of

externalism set out by Putnam and Burge respectively are suitable models for the position that I have been trying to develop, and I will then consider some of the options open to the physicalist to reconcile the externalistic individuation of thoughts with the identity claim. I will conclude with some observations on the problem of reconciling the externalistic individuation of thoughts with self-knowledge.

2.The externalistic individuation of thoughts

2.1a.Putnam 's Twin Earth

Putnam (1975c) asks us to suppose that there is a planet called twin earth where the seas, rivers and lakes are filled with a substance which superficially resembles water, but whose chemical structure is XYZ, rather than H2O. The word 'water' is used on earth to refer to the substance whose chemical structure is H2O, whereas on twin earth it is used to refer to the substance whose chemical structure is XYZ. Suppose that Oscar is an inhabitant of earth, who has a molecule for molecule identical twin on twin earth, and who says things like 'I think there is some water two miles along the track', and 'I would like a glass of water'. Suppose also that Oscar expresses his thoughts and desires using the same words that his twin on twin earth uses to express his thoughts and desires. Putnam argues that an individual can only have the thought that there is a glass of water in front of him, if in actual fact he has the capacity to refer to water, which is that substance with the chemical structure H2O. This seems unobjectionable; but the consequence of accepting this point is that we are forced to deny that Oscar and his twin can be having the same thoughts when they are staring at the glass of water in front of them, despite the fact that they are indiscernible in all internal physical and psychological respects.

Putnam 's point is that having thoughts about water presupposes having the capacity to refer to water, but that having the capacity to refer to water presupposes that one is standing in some direct or indirect causal relation to that substance which has the chemical structure H2O. The actual chemical structure of the substance referred to is argued to be relevant to fixing the meaning of the word 'water', so it appears to follow that Oscar and his twin are having different thoughts as a result of the physical differences in their respective environments. The statement, 'Oscar thinks that there is a glass of water in front of him', is therefore not just a statement about what is going on inside Oscar 's head, but is in part a

statement about the nature of the environment that he inhabits. In order to deal with the differences in content, which arise as a consequence of the differences in the physical environment, Putnam draws a distinction between psychological states in a wide sense, and psychological states in a narrow sense (1975c: 220).

Oscar's thought that there is a glass of water in front of him is a psychological state in the wide sense. Oscar can only be said to have such a thought if has the capacity to refer to that substance whose chemical structure is H2O, and this presupposes that Oscar is standing in some direct or indirect causal relation to that substance in his environment. Oscar's twin cannot be said to have the same thought, since he cannot be said to have the capacity to refer to that substance whose chemical structure is H2O. It might be said that Oscar and his twin are in the same psychological state in the narrow sense in that they are molecule for molecule identical, and furthermore, Oscar and his twin would probably behave in much the same way whenever they had the thoughts expressed by the words 'there is a glass of water in front of me', or 'there is some water two miles along the track'. But Oscar and his twin cannot be said to be in the same psychological state in the wide sense. Oscar's twin did not acquire his concepts in an environment in which the word 'water' refers to H2O, so he does not have the capacity to refer to that substance which Oscar refers to in expressing his thoughts and beliefs about water. The upshot of Putnam's twin earth thought experiment is that which thoughts an individual can be said to have seems to be dependent on the nature of his physical environment, and on the nature of the causal-historical relationship that he bears to his environment.

2.1b.Burge 's Counterfactual Linguistic Communities

Burge (1979) asks us to suppose that a patient, Bert, visits his doctor with the complaint that his arthritis has spread to his thigh. The doctor assures Bert that arthritis is a disease which can only affect the joints, so whatever is causing the discomfort, it cannot be arthritis. Burge argues that individuals use such terms deferentially to the experts in their linguistic community, and because of this Bert is prepared to take the doctor's word for it. Burge then asks us to suppose that there is a counterfactual situation in which Bert's internal physical properties remain fixed, but in which there is a difference in the use of the term 'arthritis'. In the counterfactual situation, the term is used by the experts to refer to a disease which is not restricted to the joints. In both the actual, and in the counterfactual linguistic community, Bert

has the thought expressed by the words, 'my arthritis has spread to my thigh'. But despite the fact that Bert's internal physical properties are held constant, the thought which Bert expresses in the actual linguistic community will be different from the thought which Bert expresses in the counterfactual linguistic community.

Burge suggests that although we would be prepared to attribute a false belief about arthritis to Bert in the actual linguistic community, we would simply not attribute a belief about what we call 'arthritis' to Bert in the counterfactual community. Burge insists that it is an important part in his argument that we commonly attribute thoughts to individuals despite their lacking a complete mastery of some notion in the content of their thoughts, which we can do only because we know that the individual uses his concepts deferentially to the experts in his linguistic community. Burge's conclusion is that a difference in an individual's linguistic community will be enough to constitute a difference in the content of his thoughts, even if there is no difference in the individual's internal physical and psychological properties. This clearly requires that we accept the claim that what an individual means by his words is what the words mean in his linguistic community, otherwise there will be no reason for saying that Bert's thoughts differed solely in virtue of the difference in the linguistic community to which he belonged. So the upshot of Burge's thought experiment is that which thoughts an individual can be said to have seems to be dependent on the nature of his linguistic community, and on the nature of the causal-historical relation that he bears to his community.

2.2.Strengthening The Arguments By Weakening Them

It would be nice if I could let matters rest with Putnam and Burge, but I am inclined to suspect that the arguments they advance in support of the externalistic individuation of thoughts are questionable. Or at any rate, they are not suitable models for the position that I have been trying to develop. The central problem with these arguments is that their attempt to motivate externalism seems to compromise the individual's awareness of certain aspects of his own thoughts. There is something not quite right about attributing thoughts to individuals who do not fully understand the implications of having them, unless it is clear from the start that the thoughts have been appropriately qualified to take this into account. This is particularly evident with regard to Putnam's version of externalism, which is very much weakened by his failure to appreciate the significance of the agreement there would most likely have been between Oscar and his twin, if they had been asked what they were thinking

about. Putnam's position rests on the assumption that the meaning of the word 'water' depends on the nature of the physical environment in which it was learned. If this is correct, it will have a bearing on the content of the thoughts attributed to inhabitants of different physical environments. Oscar and his twin will necessarily express different thoughts, even though they appear not to. But as Glock and Preston (1995: 519) point out, Putnam's arguments distort the conceptual connection between meaning on the one hand, and explanations of meaning on the other. Just as the meaning of a word cannot be severed from an explanation of its meaning, what an individual means when he is using a word cannot be severed from the explanation that he gives of its meaning. So if Oscar and his twin were to agree in the explanations they gave of their thoughts, it should follow that both Oscar and his twin were thinking the same thoughts when they claimed to be thirsty. The difference in the chemical structure of water on their different planets would make no difference to what they were thinking, unless they indicated their awareness of this difference in the explanations they gave.

Similarly, the problem with Burge's position is that it over-estimates the extent to which the individual's participation in his linguistic community is relevant to what he means. Or rather, Burge seems to have failed to allow sufficient scope for the individual to diverge from the linguistic community in particular situations, such as Bert's visit to the doctor. In these cases, we are more inclined to give credence to what the individual tells us he is thinking, rather than what the experts in his linguistic community dictate that he is thinking. Bert is said to be thinking about arthritis only because we attribute to him this concept despite his incomplete understanding of it, but in so doing we would be inclined to qualify this attribution by saying that Bert does not know that arthritis can only affect the joints. What we are inclined to say about Bert falls short of what we are inclined to say about the experts who have a more complete understanding; so in this respect, it looks as if considering Bert in the counterfactual situation cannot really force us to attribute different thoughts to him, despite the fact that the meaning of the word 'arthritis' is somewhat different. Unless Bert were to offer different explanations of what he was thinking in the different situations, we would not be inclined to attribute different thoughts to him, regardless of the different meanings of the terms in his linguistic community.

What needs to be worked out here is a way of making the point that what an individual thinks can be dependent on factors external to his body, without creating the kind of tension found

in Putnam and Burge. It seems to me that the most plausible way to make this point is to recognise the extent to which an individual's possession of his conceptual skills is dependent on his involvement with other people, whilst at the same time refusing to credit the individual with conceptions of his situation which are too sophisticated to cohere with his actual competence in exercising his skills. This will secure the dependence of thought on factors outside the individual's body, but it will do it in such a way as to ensure that what the individual thinks is not dependent on factors of which he might be ignorant. So although Burge is correct to insist that the individual can only be said to have his conceptual skills through participating in a particular linguistic community, our attributions of thoughts to certain individuals must be qualified in order to avoid seeming to credit them with an awareness of factors of which they are in fact ignorant.

Here is my own example. Suppose that an individual picks up his car from the garage after having it serviced. His invoice details that his sump has a hair-line fracture in it, although it has not been attended to as part of the service agreement. It would be rather misleading to attribute the thought that the sump is fractured to the individual, without qualifying it in the appropriate manner, if he then set out from the garage on a lengthy journey without taking extra oil along with him. Not qualifying this attribution would seem to credit the individual with an awareness of the implications that the fractured sump has for the lubrication of his engine, whilst his subsequent behaviour would seem to indicate that he did not fully understand what it means to have a fractured sump. The point is that what an individual can be said to think must be constrained by his level of conceptual competence as it is displayed in the various ways in which he would respond to the demands of the situation in which he finds himself, for in attributing thoughts to individuals we are expressing an interest in the extent to which they are responsive to these demands. Here we are interested in the extent to which the fractured sump figures in the individual's thinking as presenting him with a reason for taking the appropriate precautionary measures to prevent his engine seizing up. An individual who claims to have the thought that his sump is fractured is expected to display an awareness of the implications that it has for the performance of the engine; but if the individual fails to take any of the precautionary measures that would normally be expected in such a case, what he can be said to think would have to be qualified accordingly.

It now looks as if the argument can be strengthened in such a way as to keep the externalist's central intuition in place, without compromising the individual's own understanding of his

thoughts. This might be achieved through appealing to what McDowell (1986) refers to as object-dependent thoughts. Object-dependent thoughts are thoughts which are expressed through the use of demonstratives, which set up a logical connection between these thoughts and the objects they are about. The thought expressed by the words '*that* cat is about to jump onto the table', is an object-dependent thought; it is logically connected to the cat itself, in such a way that the individual would not have been able to have precisely that thought if the cat had not been there. Presumably, the individual's internal physical properties could be held constant whilst he is considered in a counterfactual situation in which there is no cat present. In the counterfactual situation, the individual cannot be said to have this same thought, since the thought is logically tied to the cat itself; so it follows that whether or not the individual has this thought about the cat is determined independently of determining whether he has the same internal physical properties in these different situations. It does not make any difference whether he has the same internal physical properties or not; what makes the difference is the presence of the cat.

This seems to be a striking illustration of the way in which thoughts are dependent on factors external to the individual, to the extent that their alleged dependence on the individual's internal physical properties is compromised, but which does not force us to attribute thoughts to individuals that outstrip their actual level of conceptual competence. It is crucial to realise that object-dependent thoughts are intrinsically related to the objects they are about, in that the object itself is said to figure as a constituent in the individual's thoughts; yet at the same time, object-dependent thoughts are nonetheless individuated according to the individual's conception of these objects, rather than by the perhaps unknown properties of the objects themselves. What the individual thinks when he has the thought that that cat is about to jump onto the table is individuated according to the individual's own conception of the cat, since the cat itself can only be said to figure in the individual's thoughts in the first place, if it can do so in such a way as to respect the individual's level of conceptual competence. This same point can be extended to cover non-demonstrative cases: the fractured sump can be said to figure in the individual's thoughts as presenting him with a reason for taking extra oil on his journey. Again, if this is the case, the individual must understand the implications of having a fractured sump for the lubrication of the engine. For although the fractured sump itself can be said to figure in the individual's thoughts as presenting him with a reason for acting to avoid engine seizure, it cannot do so in such a way as to outstrip his actual level of conceptual competence. But if it is correct to say that objects can figure constitutively in an individual's

thinking, in these cases presenting him with a reason for acting in a certain way,[1] then the externalistic individuation of thoughts can be said to undermine the important physicalist assumption that an individual's mental states and events are identical with physical states and events of his brain.

2.3.An Objection To Object-Dependency

Noonan (1993) disagrees that object-dependent thoughts pose such a problem for the identity claim. By Noonan's lights, the dependency which appears to threaten the identity claim is only superficial; it is not to be taken as signifying the non-internality of the mental in the stronger sense that I am looking for. Instead, object dependent thoughts should more properly be construed as complexes, consisting of internal psychological states plus contingently existing external factors, and as such the dependency in these thoughts should not be taken to suggest the non-internality of psychological states and events. The crux of Noonan's objection is that so-called object dependent thoughts are redundant in the explanation of an individual's actions, and this makes it reasonable to suppose that they are not, properly speaking, psychological states. Here is why. Suppose, first, that an individual kicks the cat, and that the individual's action is explained by citing the object dependent thought, 'he thinks that that cat is about to attack him'. Now suppose a counterfactual case in which everything remains unchanged apart from the fact that the individual is hallucinating the cat, and that he lashes out into thin air. The individual cannot be credited with having the object dependent thought in the hallucinatory case, for the simple reason that the non-existence of the cat is sufficient to defeat this ascription, yet the individual's behaviour is identical in both cases. But since the individual's behaviour is identical in both cases, and since the object dependent thought cannot be attributed to the individual in the hallucinatory case, it seems to follow that the individual's behaviour can be adequately explained in each case without having to cite any object dependent thoughts at all. The content of the individual's psychological states in the hallucinatory case represents a proper subset of the content of his psychological states in the veridical case, and it looks as if that subset is sufficient to explain the individual's behaviour in both. On these grounds, Noonan suggests that object dependent thoughts are in

[1] In saying that objects figure constitutively in our thinking as presenting us with reasons for acting, I do not intend to imply that this is the only way in which objects can figure in our thinking. Unless we are ignorant of the sump's function, the fractured sump at the same time figures in our thinking as the cause of engine seizure, and the cat which figures in our thinking as a reason for removing the bottle of milk from the table at the same time figures in our thinking as the cause of the smashed bottle of milk the day before.

fact redundant in any adequate psychological explanation of an individual's behaviour, and that object dependent thoughts should rather be construed as complexes:

> if a subset of (so-called) psychological states is demonstrated to be redundant in the psychological explanation of action, this is surely reason to regard them as not, properly speaking, psychological states at all (like knowledge, which is best regarded not as a psychological state, but as a complex consisting of a psychological state (belief) plus certain external factors- not because its status as knowledge is causally irrelevant in action explanation, but because it does not have to be cited, as such, in the psychological explanation of action at all). (1993: 291-2).

The problem with this objection is that it trades on the idea that the explanation of the individual's actions will be the same in both cases, which I doubt. In the hallucinatory case, we would cite the individual's thought that that cat is about to attack him; but the explanation would be misleading and incomplete, unless we expanded on it by saying that he was hallucinating. Or in other words, we would not have given an adequate explanation of the individual's behaviour if we simply said that he thinks that that cat is about to attack him; we would rather say that he is having a hallucination of a cat that is about to attack him. But no such incompleteness affects the explanation in the veridical case, since citing this thought without qualification is sufficient to explain why the individual is acting as he is. To the individual, there is perhaps no difference in each case, but the difference in the external circumstances is registered in the different explanations *we* would give of his behaviour.

Noonan's objection is based on the assumption that the content of the individual's psychological states in the hallucinatory case represents a subset of the content of his states in the veridical case. But if this is to show that the 'extra' content in the veridical case is redundant in *our* explanation of the individual's behaviour, it would have to be the case that the explanations were identical. It seems to me that this is not so, given that the non-existence of the cat in the hallucinatory case would be a relevant factor in our explanation of the individual's lashing into thin air. It may be that his *bodily movements* can be explained in the same way in both cases, and it may be that the individual would explain why he was acting as he was in the same way in both cases, but that does not mean that we would explain why he was acting as he was in the same way in both cases. As such, it seems to me that Noonan's argument is inconclusive. It seems to me that there are reasonable grounds for saying that the externalistic individuation of thoughts is incompatible with the identity claim, more so if we can make use of the idea of object dependency. For this idea implies that the individual's thinking is a unitary act which incorporates the object constitutively, which is to say that his thinking about the cat cannot be identical with events in his brain. But before concluding that

the externalistic individuation of thoughts is not compatible with the identity claim, it is worth noticing that there are a number of arguments designed to establish their compatibility, to which I now turn.

3.Reconciling externalism with the identity claim

3.1.The Dual Component Theory

One way of reconciling externalism with the identity claim is to exploit Putnam's distinction between psychological states in a wide sense and psychological states in a narrow sense. The externalistic individuation of thoughts forces us to recognise that there is a method of individuation which leads to the attribution of thoughts to individuals on the basis of the relations in which they stand to their environment. The availability of this method of individuation seems to compromise the principle that mental states and events are identical with physical states and events in an individual's brain, and the principle that indiscernibility with respect to physical properties guarantees indiscernibility with respect to mental properties. The externalistic individuation of thoughts ought to force us to abandon these principles, since there is no longer a guarantee that any two individuals who are alike in all physical respects will be alike in all mental respects. But if there is an alternative method of individuation available, which respects the physicalist's principles, and which can co-exist with the externalistic method of individuation, it looks as if externalism will be reconcilable with the identity claim after all.

Fodor (1992) argues that the twin earth thought experiments simply serve to highlight the applicability of two different methods of individuation to individuals, by highlighting the fact that our attributions of thoughts to individuals are likely to come into conflict according as we focus on one method of individuation rather than the other.[2] Fodor's suggestion seems to be that the different methods of individuation generate conflicting attributions of thoughts, but that the mere possibility of such a conflict poses no threat to physicalism. It simply underlines the fact that if we adopt a method of individuation which attributes thoughts on the

---

[2] In fact, Fodor (1992: 666) thinks that it is unlikely that we can have a method of individuation which bases its attributions of thoughts on an individual's relations to his environment. Although he does not deny that there are such relations, he is keen to emphasise that we cannot make a science out of them. He thinks that we are only ever likely to have a method of individuation which bases its attributions on the individual's internal physical properties, in such a way as to respect the supervenience thesis, since it is only these properties which will turn out to be generalisable in a science of psychology.

basis of the individual's relations to his environment, there will be the risk that the thoughts attributed in this manner might not be the thoughts that would have been attributed to him on the basis of his internal physical properties. Fodor's idea is that since it is only thoughts attributed on the basis of the latter properties which figure in our psychological theories, or in our rationalising explanations, it is only psychological states in a narrow sense which are relevant to the explanation of an individual's behaviour in terms of what he thinks and believes; it is only psychological states in a narrow sense which tell us what the individual had in mind, and it is what the individual had in mind that caused his behaviour (1992: 659).

If Fodor is right, the availability of an externalistic method of individuation need not compromise the physicalist's identity claim. We are only forced to see that the psychologist is obliged to adopt that method of individuation which picks out psychological states according to his interests in causally explaining an individual's behaviour, and that he must therefore adopt that method of individuation which permits the attribution of thoughts on the basis of the individual's internal physical or neurological properties. This means that Oscar and his twin must be considered to be in the same psychological state in so far as our interest in attributing thoughts to them is to explain their identical behaviour of reaching for the glass of water in front of them. The psychologist must therefore restrict his investigative domain in such a way that only psychological states in the narrow sense will be allowed to figure in his generalisations, and this restriction is effected by adopting the method of individuation which attributes psychological states on the basis of the individual's internal physical properties. This restriction will then successfully avert the threat which the thought experiment appeared to generate, for there is now no reason to insist that the psychological states which figure in rationalising explanations cannot be identical with those physical or neurological states that causally explain the individual's behaviour.

Colin McGinn (1982) also defends this response to the problems posed by the externalistic individuation of thoughts, by pointing out that psychological states have two separable components, which correspond more or less to Putnam's wide and narrow states. Psychological states have a cognitive component, which is relevant to the rationalising explanations we give of an individual's behaviour, and a referential component, which is relevant only in so far as we are interested in these states as representations of some features of the individual's environment. McGinn's suggestion is that the cognitive components of our psychological states are characterised by the properties which are constitutive of their causal

role, whereas the referential components are characterised by those properties which are constitutive of their truth-conditional relations to features in the individual's environment which they represent. With respect to the twin earth thought experiment, McGinn's suggestion is that the reason we get conflicting standards of individuation, according as we concentrate on the cognitive or referential component of psychological states, is this: content is a hybrid of conceptually disparate elements, both of which inform our concept of thought and belief, but depending on what our interests are, we tend to let one component of content eclipse the other. Conflict is only natural therefore, given that our psychological states have this dual nature; and it is only with respect to their cognitive component, which is causally explanatory of an individual's behaviour, that psychological states need be regarded as falling under the physicalist's principles (1982: 214-216).

The dual component theory of thought might appear to explain the conflict which the twin earth thought experiments bring to light, and it might appear to offer a reason for denying that externalism poses a serious threat to the identity claim; so, it might also appear that the dual component theory of thought rescues the possibility of identifying mental and physical states and events, and hence of giving a causal account of rationalising explanations. But it seems to me that the dual component theory only looks to be successful because it feeds on a particular weakness in the original argument that Putnam put forward. The twin earth thought experiment seemed to demand a way of insulating an aspect of Oscar and his twin's thoughts from the environmental factors which were said to be relevant to the individuation of their content. It is only in terms of such an insulated aspect of their thoughts that we would have been able to rationalise their identical behaviour in the same way, despite the fact that the externalistic method of individuation forced us to admit that Oscar and his twin were thinking different thoughts. The dual component view appears to secure this insulated aspect by postulating a separable cognitive component, which not only explains why the twins acted in the same way given its immunity to the environmental factors which the externalistic method of individuation emphasises, but which is at the same time suitable for identification with an internal physical or neurological state of their brain. But despite its apparent attractions for the physicalist's identity claim, it is questionable whether the dual component theory of thought actually makes much sense.

The first thing to note is that there does not seem to be anything positive that can be said about the separable cognitive component of these dual component states, other than the fact

that they play a causal explanatory role with respect to the individual's behaviour. The internal cognitive component is supposed to be that component which is cited in rationalising explanations of the individual's behaviour, but given that it is also supposed to be non-semantically characterised, it is difficult to see how it can be a *cognitive* component at all. Presumably, if anything is to count as a cognitive component of a psychological state, it must be possible to specify what that component is supposed to be cognitive of; but in the dual component theory, the component which purports to fill this cognitive role is said to be devoid of those properties which are constitutive of its referential relations to the individual's environment. So what seems to underlie this attempted reconciliation of externalism with physicalism is the assumption that we can separate an individual's cognitions, which figure in rationalising explanations of his behaviour, from the fact that they are cognitions *of* the individual's environment, in which he is so behaving.[3]

Fodor and McGinn are both clearly right to insist that it is what the individual has in mind that explains his behaviour, but there is no good reason to suppose that what a person has in mind must be an insulated component of a psychological state, other than the fact that it provides the physicalist with an internal component that can be identified with a physical state of the brain. There seems to be a slippage here, which the physicalist's terminology might have encouraged, between what an individual has in mind, in the sense of what he is thinking about, and what he has in mind, in the sense of the physical state of his brain which is causally explanatory of his bodily movements. This slippage is crucial to the reconciliation of the externalistic individuation of thoughts with their causal explanatory role. But it can easily be seen to be suspect, if we refuse to agree with the particular way in which Putnam sets out his externalistic intuitions in the first place. There is no need to insulate a cognitive component of thoughts in order to explain why the twin's behaviour was identical. There is a sense in which their thoughts are *already* insulated from the differences in their environments, if the different chemical structures simply do not figure in the explanations they would give of what they are thinking. This is precisely where Putnam's argument seems to go wrong. If it is correct to say that what an individual thinks is exhaustible within the explanations he might give whenever he is asked, then it is incorrect to say that what an individual thinks can be partly fixed by the unknown chemical structure of the physical environment he inhabits. The chemical structure of the substances in his environment makes

---

[3] McDowell (1986: 160), captures this point quite nicely: "It is impossible not to be concerned about the boundary around the internal component of the two-component picture, and the darkness within it, if one is concerned at all about the relation between subjectivity and the objective world."

no difference to what the individual thinks, unless it figures somewhere in his conception of the situation; and since it is how the individual conceives the situation that determines how he will act, there is no reason to make the problematic bifurcatory move that Fodor and McGinn make in response to externalism.

3.2.Events And Their Descriptions

The central difficulty that the physicalist faces is that of reconciling the object dependency in certain types of thoughts with the identity claim. That objects can be said to figure constitutively in an individual's thinking presents the problem of explaining why the mental event of thinking that the sump is fractured can be partly constituted by the fractured sump itself, whereas the physical events occurring in the individual's brain at the same time cannot be said to be so constituted. The implication of admitting that objects can figure constitutively in an individual's thinking seems to be that the identity thesis has to be rejected, since there can be no plausible means of explaining how the physical events in the brain, with which this mental event is to be identified, can have the fractured sump as a constitutent part. No doubt this problem could have been dealt with by the dual component theory of thought, but I have already found reason to reject that approach. There is, however, an alternative way of dealing with this problem, which is simply to deny that the implication of admitting that objects can figure constitutively in an individual's thoughts is that the identity claim has to be rejected. This move attempts to reconcile the fact that certain thoughts are individuated in terms of external objects, with the fact that they are nonetheless identical with physical events which are not so identified, by taking advantage of the distinction between events and their descriptions.

Davidson (1994: 58-9) attempts to effect this reconciliation by comparing mental events and states to the state of being sunburned. Sunburn is a state which is individuated by its causal relations to the sun, an object which is external to the sunburned patch of skin itself. But this does not mean that sunburn cannot be identical with a physical state of the skin, just because it is identified as the type of burn it is in terms of its causal relations to an object outside the boundaries of that patch of skin itself. In the same way, mental events are individuated in terms of their causal relations to the objects and events that they are about, but this does not mean that mental events cannot be identical with physical events in the body. This apparently shows that there is no obstacle to holding that mental events can be identical with physical

events in the body *and* that mental events are individuated in terms of an individual's causal-historical relations to his environment; for although physical states and events in the brain do not themselves presuppose the existence of external objects, the physicalist could easily agree that some of their descriptions might. The point could be put by saying that even if some mental events are logically related to the external objects they are about, it does not follow that these mental events cannot be identical with certain physical events in the individual's brain. For although the physical events in the individual's brain cannot be said to be logically related to such external objects, the physicalist could easily reply that the logical relation in question need only be admitted to hold between certain descriptions of these events and objects, as opposed to between the events and the objects themselves. So there is simply no entailment from the claim that mental events are individuated in terms of external objects to the claim that mental events cannot be identical with physical events in the brain.

The problem I have with the sunburn analogy is that, whilst it is correct to say that the burned patch of skin is identified as a patch of sun-burned skin in virtue of its causal historical relations to the sun, it is not so clear that we can apply the same reasoning to the brain events with which mental events are supposed to be identified. The analogy looks as if it is meant to explain how non-relational brain events can be identical with mental events which are individuated in terms of relational factors, but it seems to me that it explains no such thing. Because the analogy starts out with a burned patch of skin, which is then identified as the type of burn it is in virtue of its cause, we need to start out with a thought, and then identify it as the type of thought it is in virtue of its cause. So if the analogy is to be applied in this instance at all, we would have to be able to work with the following distinction: it would have to be possible to start out with a non-relationally identified thought event, and then identify it as the type of thought it is in terms of its causal relations to objects or events in the individual's environment. But even if this distinction makes sense, it is not what Davidson wanted his analogy to explain. It does not explain how non-relationally individuated brain events can be described as relationally individuated mental events, since the analogy only seems to let us start out with non-relationally individuated mental events. In order to apply in this instance at all, there would have to be further grounds for assuming that non-relationally individuated mental events are identical with non-relationally individuated brain events. This, however, is simply to presuppose as intelligible what the analogy was supposed to help us understand in the first place.

Macdonald (1990: 400-402) develops a similar line of defence, again emphasising the distinction between events and their descriptions. In making this distinction, she draws a further distinction between constitutive properties of events and characterising properties of events. Quite simply, constitutive properties of events are their essential properties, which cannot be altered without altering the events themselves; characterising properties, on the other hand, are non-essential properties, which are instantiated in events by virtue of certain of their descriptions. The relevance of this distinction to preserving the identity claim is now fairly straightforward: intentional mental properties supervene on an individual's internal physical properties and his causal relations to his environment, but there is no obstacle to saying that the events which instance these mental properties are nonetheless events in the individual's brain. This reconciliatory move can be made without too much trouble if we agree that the latter properties of the individual's brain events are characterising rather than constitutive properties, which is to say that since they are instanced in these events only by virtue of certain of their redescriptions, the externalistic individuation of thoughts is perfectly compatible with the identity claim.

Unfortunately there does not seem to be a neutral standpoint from which to distinguish between those properties which qualify as constitutive and those which qualify as characterising. Any decision on whether one type of property rather than another is constitutive of the token event will obviously be determined by the requirements of the framework in which the event's properties are to be typed, which makes the decision to type physical properties as constitutive rather than characterising a foregone conclusion. But perhaps the physicalist is entitled to this bias, since the question is whether the object dependence in thoughts can be reconciled with the identity thesis. Suppose, then, that we grant this much for the sake of the argument. We still need to understand *how* the token event's characterising properties are related to its constitutive properties, otherwise this distinction will be of no real use, and this raises the problems already encountered with regard to Macdonald's notion of property coinstantiation. It has yet to be made clear how mental properties, which supposedly characterise an event, can be coinstantiated by physical properties, which are supposedly constitutive of the event. It has yet to be made clear how, for example, the individual's thinking about the fractured sump can be coinstantiated by particular neural events which happen to be occurring in his brain at the same time. It does not help matters to insist that these neural events instantiate the relevant mental properties because they have been caused by light emissions coming from the fractured sump, since

there would then have to be some sort of an explanation as to how these causal linkages could be characterised in intentional terms, which takes us directly back to the problem of explaining how mental and physical properties can be coinstantiated. But until these difficulties have been resolved, and it is by no means obvious what would resolve it, the distinction between constitutive and characterising properties does not explain how to reconcile the object dependence in our thoughts with the identity thesis.

4. Thinking and knowing what one is thinking

I have been appealing to the externalistic individuation of thoughts in order to put pressure on the identity claim. But there could be a problem with taking this line, for it is sometimes argued that the externalistic individuation of thoughts poses a problem for the individual's privileged knowledge of what he is thinking. The reasoning behind this objection is simple: if what an individual thinks is dependent on factors which lie beyond the confines of his body, he will not be in the position to know what he is thinking until he undertakes an empirical investigation of his environment. This seems to have the consequence that a second person will be in just as good a position to tell what an individual is thinking about as the individual himself. Worse than this, a second person might well be in a better position to tell what the individual is thinking about than the individual himself. Oscar and Bert immediately lose authority on their own thinking, once it is granted that their thoughts are individuated by factors which are equally accessible to others. To my mind, this is unacceptable. I have already suggested that part of what we are doing in attributing thoughts to an individual is expressing an interest in his awareness of the demands of the situation in which he finds himself. So it must be important to acknowledge the individual's own perspective on his situation, even where it seems to diverge in striking ways from what the experts in his linguistic community would be inclined to grant. I have been arguing that this forces us to weaken the externalist's position somewhat, but Burge insists that his stronger position can easily accommodate the individual's authority on his own thoughts.

Burge (1994) defends his stronger position by arguing that whilst the enabling conditions for having a particular thought must be presupposed in the thinking of that thought, the individual can be said to know what he is thinking without having to know that these enabling conditions actually obtain. The individual can be said to know what he is thinking in so far as he has the capacity to think the first order thought self ascriptively. An enabling condition for

having the thought that there is some water two miles along the track is that the individual is standing in some complex causal relation to water; an enabling condition for having the thought that his arthritis has spread to his thigh is that the individual is situated in a linguistic community in which the term 'arthritis' refers to a disease that is not restricted to the joints. So the logical presupposition of having the thought about water is that the individual is standing in some complex causal relation to the substance with chemical structure H2O, and the logical presupposition of having the thought about arthritis is that he is situated in the appropriate linguistic community. Burge argues that although these enabling conditions must obtain if the individual is to have these thoughts, it does not follow that the individual cannot know what he is thinking unless he knows that the enabling conditions do in fact obtain. What is required is simply that the individual thinks these first order thoughts whilst exercising his second order self ascriptive skills:

> To think of something as water, for example, one must be in some causal relation to water- or at least in some causal relation to other particular substances that enable one to theorize accurately about water. In the normal case, one sees and touches water. Such relations illustrate the sort of conditions that make possible thinking of something as water. To know that such conditions obtain, one must rely on empirical methods...But to think that water is a liquid, one need not know the complex conditions that must obtain if one is to think that thought. Such conditions must only be presupposed. (1994: 69-70).

> One knows one's thought to be what it is simply by thinking it while exercising second-order self-ascriptive powers. One has no 'criterion', or test, or procedure for identifying the thought, and one need not exercise comparisons between it and other thoughts in order to know it as the thought one is thinking. Getting the 'right' one is simply a matter of thinking the thought in the relevant reflexive way. (1994: 72).

It seems to me that this defence can only have limited success, given that Burge's original argument for externalism can be seen to depend on the assumption that we are bound to give an individual's words the meanings they have in his linguistic community. This assumption seems to rule out the possibility of the individual diverging in what he means on a particular occasion from what the words standardly mean in his linguistic community, and at the same time it seems to commit us to the assumption that we are sometimes bound to attribute thoughts to an individual which he does not think he has. It could happen that the individual claims to be having a thought about water, when in fact the enabling conditions for having this first order empirical thought do not obtain. This could happen because although the individual is exercising his second order self ascriptive skills, the content of the first order thought over which these skills are exercised is determined by conditions which are unknown to him. The individual thinks he is having a first order thought about water; but this is defeated by the fact that the enabling conditions for having that thought fail to obtain. So

although the individual thinks that he is having a first order thought about water, his ignorance of the enabling conditions blinds him to the fact that he is having a first order thought about some other substance. As such, it does not make any difference that the second order thought takes the first order thought as its content, if the content of the first order thought can be determined by factors of which the individual is ignorant. Or, as Georgalis puts the complaint:

> If one is not aware of content in the first case, or if the causal relation determining content in the first case, does not explain the subject's awareness of content, why should content within content make a difference relative to awareness of first content? *It must be included in a way that is irrelevant to the subject 's awareness of his contents.* (1990: 106).

And the same complaint is made by Brueckner:

> *Thinking*...that I am thinking that such and such does not necessarily amount to *knowing* that I am thinking that such and such, even if it is true that I am thinking that such and such. So the question remains open as to whether my introspective thought that I am thinking that some water is dripping...amounts to *knowledge* that I am thinking that some water is dripping, given my lack of knowledge of the crucial content-determining circumstances...my knowing that I am thinking that some water is dripping requires that I know that I am not thinking that some twater is dripping. How can I know *that* if I lack knowledge of external content-determining circumstances? (1990: 449-450).

Fortunately, this same type of objection does not really affect the weaker position that I have been trying to develop. The very fact that our attributions of thoughts to certain individuals have to be qualified according to their conceptual competence is indicative of the fact that first person authority is taken for granted. It is a logical presupposition of having the thought that there is some water two miles along the track that the individual has the conceptual skills which are exercised in the expression of that thought. But it is important to realise that what the individual can be said to think on any particular occasion must be consistent with the level of his conceptual competence. This ought to be enough to rule out the difficulties affecting Burge's position: what an individual can be said to think on a particular occasion is not logically separable from his ability to explain what he is thinking on that occasion, so it follows that the criteria for attributing a thought to an individual are at the same time criteria for excluding the possibility that he might not know what he is thinking. If what the individual can be said to think is exhaustible within the explanations that he would give if he were asked, then in general there can be no aspects of his thoughts of which he might be ignorant. There might be aspects of the individual's thoughts which he finds difficult to express, but that has probably got more to do with his not having a clearly formulated idea in

the first place, or with his distressed state of mind, than with his not having access to the external factors which purportedly determine what he is thinking.

Having said that, I do think that there are some cases in which it is not so obvious that the individual has privileged access to his thoughts. Self-deception comes immediately to mind. But as a concrete example, consider this case. Suppose that an individual is on safari. He suddenly notices a snake sliding into a bush just ahead of him, and he thinks to himself, '*that* snake is venomous'. A *different* snake slides back out a second later. The individual fails to notice this difference, and he is still thinking to himself, '*that* snake is venomous'. The thought at the later moment is logically tied to the different snake, by the use of the demonstrative in the expression of the thought, although no such difference is registered by the individual. Someone walking behind the bush can see two identical snakes sliding back and forward, and he later asks the individual which snake he was thinking about. Since the individual did not distinguish between the two snakes, it seems to follow that although he was thinking about the different snakes at different moments, the individual is not able to say which snake he was thinking about at any particular moment.

No doubt more cases like this one can be imagined, but I do not think they pose any serious threat to the more central cases in which no such doubts arise. Given that we are sometimes susceptible to illusions or self-deceptions, and that we do not always have the clearest perception of things, it is only to be expected that such cases exist. But what is distinctive of these cases is that an explanation of why the individual's authority is in question is called for, whereas in normal everyday situations we do not require an explanation of why the individual's claim to know what he is thinking is taken for granted. The very fact that we want an explanation as to why the individual fails to have authority on his thinking in certain situations is indicative of the fact that they are atypical and non-ordinary, and that in assessing their impact on the issue of self knowledge we have to treat them on their own merits as special cases.

5.Conclusion

I have been trying to motivate a version of externalism that is weak enough to preserve the individual's authority on his own thinking, but which is strong enough to put pressure on the idea that mental states and events are identical with physical states and events. Even if the

argument that I have put forward does not conclusively refute the physicalist's position, it does suggest reasons for doubting that the identity claim can be preserved unproblematically. Certainly, in the case of object-dependent thoughts, the argument that I have been developing seems to be much stronger. If factors in the individual's situation can be said to figure constitutively in his thinking, as presenting him with a reason for acting in a certain way, then it is very difficult to retain the idea that mental events are identical with internal behaviour-causing physical events in the individual's body. Further, if what an individual thinks can be said to be logically dependent on the presence of the object itself, then what he is thinking can be said to differ according to the layout of his situation, rather than according to the internal physical properties instantiated by his neural events. So if it is correct to say that objects themselves can figure constitutively in our thinking, then it makes it rather unclear that the identity claim provides us with the correct expression of the relation between the mental and the physical, and hence also that the causal account of rationalising explanations provides us with the correct account of how reasons explain actions, which is exactly the conclusions I have been looking for.

Chapter 7: Thinking and the Brain

1.Introduction

The approach to thinking that I have been developing focuses on the individual's involvement in human relationships. I have argued that it is only if the individual can be said to have an awareness of the demands of human relationships that he can be said to have the capacity to think. But it might seem that the result of placing the emphasis on the individual's involvement in human relationships as the crucial factor is that the neurobiological functioning of the brain is in danger of being completely ruled out of consideration. For if what it means to say that an individual has the capacity to think is intrinsically connected to what it means to say that he has the capacity to be involved in human relationships, then it is not obvious what the brain has got to do with thinking. I have indeed said very little about the brain and its relevance to my position, which might not be surprising, given that I have been concerned to develop a non-physicalistic approach to thinking. But it might seem that this is not justified, since I have in effect disregarded the hard scientific facts of the matter in order to push the thesis that mentalistic explanations are autonomous with respect to physicalistic explanations in the deep sense. Admittedly, the hard scientific facts of the matter *have* been ignored, and perhaps justifiably so. For it is not very clear what the discovery of such facts, whatever they are, could tell us about what we mean when we say that an individual has the capacity to think; nor, for that matter, is it very clear that the discovery of these facts would force us to say that mentalistic explanations must implicate physicalistic explanations in carrying out their work.

Certainly, it would be ill-advised of me to ignore the brain completely, even if the point of considering the brain in its relation to thinking is only to offer reasons for the mistrust I feel, regarding the confidence with which some philosophers assert that thinking goes on in the brain, or that the brain is complex enough to be our thinking thing.[1] It is important to note

---

[1] Searle (1991: 50) writes: "It's an obvious fact that the brain has a level of real psychological information processes. To repeat, people actually think, and thinking goes on in their brains." And in a similar vein, Flanagan (1992: 60), writes: "The brain is a supremely well connected system of processors capable of more distinct states, by several orders of magnitude, than any system ever known. This, I hope, provides some reassurance that the brain is complex enough to be our *res cogitans*- our thinking thing." Both of these claims fall foul of what Kenny (1985), refers to as the 'homunculus fallacy', which is the mistake of applying mental concepts to parts of human beings; in this case, the mistake is to apply the concept thinking to the brain, when its grammar permits only an

that, although it is not immediately obvious what importance I can attach to the brain, the approach to thinking that I have been developing does not commit me to denying that it has any role to play at all. There is no denying that the functioning of the brain, and the rest of the nervous system for that matter, must figure in physicalistic explanations of human behaviour. That much seems obvious. What does not seem obvious is that this should force us to postulate a relation between the mental and the physical that would be consistent with the metaphysical unifying principles inherent in the non-reductivist's metaphysical framework. Resisting this particular move, however, does not force me to deny that the functioning of the brain is relevant. Without the stable functioning of the brain, the individual's behavioural possibilities and bodily capacities in general would be in danger of being severely restricted and impaired, preventing him from carrying on with his life in a normal manner. So this clearly suggests that the functioning of the brain must be presupposed at some point in my position, even though I am not prepared to elevate it to the prime status it is sometimes accorded. I will begin in the next section with a consideration of two different ways of articulating the relation that holds between thinking and the brain, one direct and unmediated, the other indirect and mediated, and I will argue that neither position is in fact satisfactory.

2.Thinking: biological or computational?

2.1a.Biological Naturalism And The Brain In A Vat Fantasy

Searle (1983) develops an account of thinking which accords the brain the type of importance I am concerned to resist. He puts forward the view that thinking is a natural biological phenomenon, in the very same sense as digestion, mitosis and photosynthesis are natural biological phenomena, which occurs or goes on in the individual's brain. According to this position, which is unsurprisingly named biological naturalism, thinking is both caused by, and realised in, the neurobiological processes going on in the brain. The relation between thinking and the brain is thus direct and unmediated: thinking is nothing over and above neurobiological processes going on in the brain. Searle attempts to motivate this conception of the relation between thinking and the brain by appeal to the brain in a vat fantasy, which is meant to show that an individual could have had the capacity to think, even though he did not

---

application to the whole human being. Whilst this is certainly correct, I would emphasise that the problem lies with the application of mental concepts to that which fails to be a person in any sense of the word, to that which fails to have the capacity to relate to other people, to that which fails to have interests, to that which lacks moral status, and so on.

actually stand in any type of relationships to other people, or indeed to his environment. The underlying assumption, which I find unacceptable, is that the individual's relationships with other people, and his relations to other things, could be dispensed with, without this making any difference at all to his thinking. Once the individual's thinking is concentrated in the brain in this manner,[2] it becomes irrelevant that he is actually involved in relationships with other people, that he is actually living through situations in which he is presented with reasons for acting in certain ways, and feeling certain things.

Yet Searle (1983: 143) points out that having mental capacities depends on having an array of practical skills and abilities, pre-intentional stances and attitudes, behavioural habits and dispositions. To my mind, this is to suggest that thinking is inseparable from the life that the individual lives in the world with other people, and that other people and other things are therefore indispensable in a way that would undermine the intelligibility of the brain in a vat fantasy. But Searle evades this conclusion by claiming that all this necessary background is similarly realised in the individual's brain and his nervous system. The immediate difficulty with this idea, however, is that whilst the individual's thinking is recognised as being inseparable from the life he lives, the life he lives is so strongly concentrated into the brain that it too turns out to be exclusively dependent on his neurobiological functioning. But this is to suppose that there is nothing more to the fact that the individual has a life to live than the fact that human life has a neurobiological dimension. By this account, if it were possible to completely reproduce the individual's neurobiology in a laboratory, and to sustain an identical level of electrochemical stimulation, this would be sufficient to reproduce the life that the individual lives. But it seems to me that regardless of how complete the replication turned out to be, even if it were molecule-for-molecule exact, the fact that the individual has a life to live could not be reproduced in this manner.

One of the difficulties with the brain in a vat fantasy is that the individual who has the capacity to think is an individual who has a conceptually constituted orientation in the world, which is shaped and constrained through his involvement in various types of relationships with other people. The fact that the individual has this orientation in the world cannot be

---

[2] The metaphor of 'concentration' is suggested by Cockburn (1985), who refers to the mind-brain identity theory as concentrating the human being into the brain. The point, I presume, is that the concept 'mind' is inextricably bound up with the concept 'person', and that talk of the mind is therefore talk of attributes, capacities, qualities, and so on, which are attributable only to living persons. 'Concentrating' the human being in the brain, severing the internal connections with the mind and the human form, is the logical consequence of identifying the mind with the brain.

reproduced by electrochemically stimulating a neurobiological replica of the individual's brain in a laboratory. The basic problem here is not that it would be technically impossible to reproduce a neurobiologically identical brain in a laboratory situation; even if that could be achieved, it would not be sufficient to reproduce the fact the individual has a life to live in the world with other people. The basic problem is rather that the individual's orientation in the world is conceptually constituted, whereas the electrochemical stimulation of the neurobiologically identical brain would be a purely causal matter. The gap between the neurobiogical dimension of the individual's life, and the life that the individual lives, is a logical gap. We would not, for instance, describe a brain in a vat as having desires or fears, as falling in love or planning to get married. Such descriptions make no literal sense when applied to the brain. So to suppose that the individual's thinking can remain inseparable from the life the individual lives, whilst at the same time be nothing over and above the neurobiological processes going on in his brain, is therefore to ignore the logical gap that separates these dimensions of human life, and to merge them together in an unacceptable manner.

2.1b.Scientific Evidence In Searle 's Defence?

It might be replied in Searle's defence that there are some indisputable scientific facts which can be cited as evidence for the claim that thinking goes on in the brain, or that thinking is nothing over and above neurobiological processes. The force of this defence, however, must be immediately weakened by the difficulties we have in making sense of the idea in the first place. It seems to me that unless we can make sense of what this claim means, it cannot strengthen its plausibility to offer evidence in favour of it; the difficulty with such a move is that before the hard scientific facts of the matter can be cited in support of the claim that thinking is realised in the brain, it has to be clear that we understand not only where to look for this evidence, but also that we understand what would qualify as evidence for such a claim. It seems to me that this presupposes that we understand what it means to say that thinking is realised in the brain, which I have already found reason to question. As an example, here is some such evidence cited by Flanagan:

> Positron emission tomography (PET) and very recently turbo-charged magnetic resonance imagining provide amazing opportunity to watch thought in action, to map experiences onto brain processes...There is robust evidence showing vivid differences in brain activity in persons engaged in phenomenologically distinct mental activities...Activity of lots of processors in different locations is required if certain kinds of thoughts are to occur. (1992: 39-40).

> Work with PET scans and other brain imaging techniques indicate that there are certain gross typological correlations between feeling and thought types and types of brain process. (1992: 47).

Although there is nothing implausible about the claim that there are vivid differences in brain activity in persons engaged in phenomenologically distinct tasks, and although it might well be the case that activity of lots of processors in different locations is required if thoughts are to occur, these facts cannot be used in support of the claim that thinking is realised in the brain, or that mental processes just are brain processes. The problem is not that the evidence underdetermines the theory; the problem is more basic than this. It is simply that without an understanding of what it could mean to say that thinking is realised in the brain, it is not very clear what kind of work the evidence is supposed to do, nor is it very clear what would count as evidence for such a claim in the first place. It seems to me that we simply do not understand what Flanagan means when he says, for instance, that brain imaging techniques provide us with an opportunity to watch thought in action, unless we already understand what it means to say that thinking is realised in the brain. How could the occurrence of brain activity, when an individual is thinking, provide support for the claim that his thinking is in fact nothing over and above this brain activity? Answering this question presupposes that we already understand the type of role that the activity of processors in the brain are required to fulfil, if they are to be understood as that which constitutes the individual's thinking. But this is precisely what we do not understand. Perhaps we only *seem* to understand what this means because we are prone to be misled by what Olafson calls the amphibious nature of these inquiries:

> what can and cannot be said in the idiom of neuroscience strictly construed seems perfectly natural to both the scientist and his audience and attracts no special attention. It is of great significance, however, that these scientific inquiries have what might be called an "amphibious" character by virtue of which they function both inside and outside the conceptual limits of their object domain as that of a physical science. This means that the understanding that informs their inquiries has sources other than the observation and analysis of what can be show to take place in the brain. Most important, that understanding informs the way in which the objects of neuroscience are conceived in the context of such inquiries. (1995: 249-50).

Olafson's point might be put by saying that it only *seems* that we can make sense of these claims, because it is difficult for the neuroscientist to completely sever the type of understanding appropriate to his domain of investigation, when he comes to study subjects in the impersonal mode, from the type of understanding appropriate to his involvement in personal relationships. The neuroscientist himself is an individual who is capable of being involved with others, including his present subjects, on a personal level. It is therefore

unsurprising that his understanding of certain aspects of the physical dimension of human life is partially informed by his understanding of the mental. This is particularly obvious when the mode of understanding the physical is brought into operation in a context in which the individual is required to carry out everyday mental activities, such as thinking about the cross-word puzzle he left unfinished on the bus, or about how awful the coffee tasted before the experiment began. So when it comes to the neuroscientist's understanding of specific aspects of the domain circumscribed by his inquiry into the physical dimension of human life, his understanding of these aspects is already influenced by his prior understanding of the everyday mental activities which his subjects are being asked to perform.

Let me pick up on another point raised by Flanagan, concerning the possibility of typological correlations between feeling and thought types and types of brain processes. The central difficulty with this suggestion is that thinking can take a variety of different forms depending on a whole host of interrelated factors, namely, an individual's experience and level of expertise, his practical abilities, his conceptual skills, and so on, and so forth. The forms that an individual's thinking can take on different occasions are various, and they fail to be type-related to the various forms exhibited by neurobiological events and processes in his brain on these occasions. The type of unity exhibited by the various forms that thinking may take is a unity deriving from the uses we make of the concept thinking itself, which we learn to apply in different cases through the course of learning to be involved in relationships with others. But this type of unity is simply not matched within the domain of neurobiology, whose principles of classification are not able to get a grip, with a view to imposing unity, on the various forms that thinking may take. This seems to be part of what Wittgenstein is trying to get at, in the following remarks:

> Is thinking a specific *organic* process of the mind, so to speak- as it were chewing and digesting in the mind? (1981: § 607)

> No supposition seems to me more natural than that there is no process in the brain correlated with...thinking; so that it would be impossible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the system continue further in the direction of the center? Why should this order not proceed, so to speak, out of chaos? (1981: § 608).

Wittgenstein's final question here may be read as expressing a rather strong point, namely, that the individual's thinking may be causally dependent on nothing more than a chaotic mass of neurobiological processes. The suggestion may be the strong point that there is no order at

all at the neurobiological level, rather than the weaker point that although there is order at this level, it is chaotic when considered from the perspective of the order constituted by mental concepts. I am inclined to interpret the remark in terms of the weaker point. The point, then, is that the order exhibited by the different forms that thinking may take is an order which is loosely imposed through the concept thinking itself, whereas the order exhibited by different types of neurobiological processes is an order which can only be in place through the concepts particular to that domain, and which is chaotic and random when viewed from the perspective of the latter. For instance, the patterns of brain processes which occur when an individual is thinking that the coffee tasted awful, or that the cross-word puzzle was quite difficult, are random and arbitrary when viewed in relation to the individual's thinking, and there is no reason to suppose that type identical patterns will be exhibited in the brains of two individuals who are thinking the same thing. This is what makes it rather implausible to assume that the order exhibited in the various forms that thinking may take can continue right through to the neurobiological processes themselves. But unless there is some way to make sense of this supposition, it leaves it quite unclear what it means to say that there are gross typological correlations between thought types and types of brain processes, or that the individual's thinking is nothing over and above the neurobiological processes going on in his brain.

2.2.Computational Processes As The Mediating Link?

Perhaps a more plausible approach is to insert some type of mediating link between thinking and the brain, in such a way as to avoid the problems which arise out of the assumption that thinking is directly realised in brain activity. This idea is explained by Fodor (1987), in his computational theory of thinking, according to which the functional relationship between computer hardware and the computational programs it runs serves as a working model for the relationship that obtains between the brain and the mental processes it subserves. The central idea seems to be that just as the computer hardware implements computational programs at an abstract functional level, so the brain can be said to implement computational processes at a similarly abstract functional level. To say that the computer hardware implements computational programs is to say that, given its abstract functional structure, it can be said to be performing computational operations on a string of syntactically structured symbols. The computational theory of thinking is therefore committed to the claim that, given the brain's

functional organisation, it can be said to be carrying out computational operations on syntactically structured symbols at a level that is inaccessible to consciousness.

The simplest way to approach the computational theory of thinking is to consider an example of the type of problem it is invoked to explain: suppose you think that Maria is pretty, and that it would be pleasant to spend some time with her; but you have heard that her father is particularly protective of her, so you decide that it would be wise to convince her father that your intentions are entirely honourable. The computational theory of thinking depends on assumptions of the following type: your being in the mental state of thinking that Maria is pretty causally brings about your being in the mental state of thinking that it would be pleasant to spend some time with her, and this mental state in turn causally brings about your being in the mental state of thinking that it would be wise to speak to Maria's father. What impresses Fodor is the fact that although the various mental states are causally connected, the generalisations of common-sense psychology nonetheless pick out these causally connected mental states by reference to the propositions they express. The causal relations that obtain between the mental state of thinking that Maria is pretty, and the mental state of thinking that it would be wise to speak to her father, somehow contrive to respect the inferential content relations that also obtain between the very same mental states, when they are picked out by the generalisations of common-sense psychology by reference to the propositions they express. The central problem is therefore to see how the generalisations of common-sense psychology, that specify causal relations between mental states, can pick out these mental states in such a manner that there are also non-arbitrary content relations between them. Fodor seems to see this as some kind of engineering problem, one that calls for the construction of a theory:

> How could the mind be so constructed that such generalisations are true of it? What sort of mechanism could have states that are both semantically and causally connected, and such that the causal connections respect the semantic ones? (1987: 14).

Fodor's solution to this problem is his computational theory of thinking: it is in virtue of the fact that there are underlying computational mechanisms that the mental state of thinking that Maria is pretty causally brings about the mental state of thinking that it would be pleasant to spend some time with her, in such a way as to preserve inferential content relations. If the underlying computational mechanisms are to explain how this can be the case, then they must be such that they bring causal and semantic relations together into the same sequence of mental states. This task appears to be exactly suited to computational mechanisms, in that

they are designed to transform one symbol into another by operating on their syntactic properties, and the transformation is effected in such a way that certain semantic relations between the propositions expressed by the symbols are preserved. There are two central points here. First, the syntax of a symbol can be thought of as an abstract feature of its shape, and the shape of the symbol can be thought of as a potential determinant of the causal role of the symbol. That is, syntax is a determinant of a symbol's causal role. Second, certain semantic relations can be "mimicked" by their syntactic relations. That is, the semantic relations which hold between two symbols, when the proposition expressed by one is semantically related to the proposition expressed by the other, are parallel to the syntactic relations in virtue of which one of the symbols is derivable from the other. So if the underlying computational mechanisms are to be effective here, their operating at this level must explain how the mental state of thinking that Maria is pretty causally brings about the semantically related mental state of thinking that it would be pleasant to spend some time with her.

One of the main difficulties I have with the computational theory of thinking is its assumption that, given the appropriate functional organisation of the brain, it can be said to implement computational operations defined over the syntactic properties of mental representations. The problem is that the computational processes are held to be implemented at a level that is inaccessible to consciousness, and the mental representations, which are postulated to provide a mechanism for getting from one mental state to another in a rational sequence, are tokened in the brain independently of the thinker's awareness of them. The computational theory of thinking thus holds that mental symbols are tokened in the brain at a level that is inaccessible to consciousness; but the problem here is that nothing can count as a symbol or a representation that is not at least available to be used in a rule governed way. This casts doubt on the idea that having a thought can be analysed in terms of having a mental symbol tokened in one's brain in a certain way, for the simple reason that such tokens cannot be said to be symbols, and nor, therefore, can they be said to express propositions. But if this is the case, it makes no sense to say that computational processes can be called upon to generate semantically coherent mental processes out of such mental representations.

There are two main reasons for this: first, if mental symbols tokened in the brain cannot be said to express propositions in the first place, then computational operations on these symbols will not generate inferential content relations between them; second, even if it is

granted that computational operations on these symbols generate sequences of causal relations that succeed in mimicking rational relations between propositions, which now ought to seem problematic anyway, this falls short of the claim that inferential content relations can actually be mechanically implemented.

A related problem is that the computational theory of thinking assumes that the postulation of underlying computational mechanisms provides a causal explanation of thinking in the sense that they constitute the mediating links between the neurobiological properties instantiated by the brain, and the common-sense psychological properties instantiated by our mental processes. Computational mechanisms are said to provide a causal explanation of thinking in that they generate rational sequences of mental states by manipulating mental representations according to their syntactic properties. But if the mechanisms operate at a level that is inaccessible to consciousness, there seems to be no more reason to say that they are rule governed than there is to say that the mental symbols they manipulate express propositions. In general, the ability to follow rules presupposes the ability to respond to the normative constraints that rules impose on one's behaviour. So if one can be said to be manipulating symbols according to their syntactic properties, one must be aware of what counts as a correct manipulation of the symbols and what counts as an incorrect manipulation of the symbols. But in the case of computational mechanisms implemented in the brain at a level that is inaccessible to consciousness, there is no way of giving content to this distinction. The most that can be said is that, given the functional organisation of the brain, some of its processes can be described *as if they were* rule governed. But this falls short of the claim that computational mechanisms actually operate in a rule governed way on the syntactic properties of mental symbols, and in turn it falls short of supporting the idea that computational processes constitute the mediating link between thinking and the brain.

3.The attitude of scientistic optimism

3.1.Seeing-Aspects Of Organisation

Part of what underlies the idea that the brain must be given such a central role in our account of thinking is the feeling that mental concepts refer to states and events whose causal efficacy with regard to human behaviour must be explained in terms of neurobiological processes in the brain. Goldfarb (1992: 112) refers to this as the attitude of scientistic optimism, the smug

and unexamined assurance that what wants explanation is obvious, and that the application of scientific tools to the problems at hand is the only approach to take. Within the context of this approach, the direction in which an answer is to be pursued to the question of the relevance of the brain to thinking has already been settled in advance: thinking is to be understood in terms of the operation of the brain; whether the latter is conceived biologically, computationally, or in some other manner, is simply a matter of detail, to be worked out once the basic orientation is in place. The difficulty which faces the approach to thinking that I have been recommending, more so when it is held up for comparison against the various scientific approaches to thinking, is that it is immediately spurned as unrigorous, unenlightened and unworthy of serious consideration. But the negative attitude that is widely expressed toward such non-physicalistic approaches is only to be expected: it is simply not on the agenda of the scientific approach to work out an understanding of the mental which would resist integration, through a battery of unifying principles, into the physicalistic framework in which its own investigations are carried out.

The battery of metaphysical unifying principles are appealed to as a means of expressing the relation that is thought to hold between the mental and the physical, when the available data are organised from the perspective of the attitude of scientistic optimism. But the connection that is subsequently postulated is simply *one* way of organising the data; it is indicative of the confidence that it must be possible for everything to be explained scientifically.[3] It might be said that this is a kind of seeing-as, which comes about through the unquestioned and unlimited application of the conceptual apparatus that gives expression to the attitude of scientistic optimism. To interpret neurobiological processes as constituting thinking is, I suspect, an achievement of the scientistic attitude, which is given expression through the unbridled application of the conceptual apparatus particular to the neurobiological dimension of human life. To see the connection between the mind and the brain as one of realisation or constitution is to have one's perceptions shaped in this way, and the approach to thinking that I have been recommending then comes to be regarded as naive and rather simplistic.

Once the attitude of scientistic optimism has been cultivated, through over-exposure to the tough-mindedness of those who stubbornly refuse to entertain the possibility that scientific tools are not immediately applicable in every instance, it becomes increasingly natural to insist that the available data can only be organised in one way, and it becomes increasingly

---

[3] Wittgenstein remarks: "Where does this idea come from? It is like a pair of glasses on our nose through which we see whatever we look at. It never occurs to us to take them off." (1967: § 103).

unnatural to return to the non-scientistic perception with which, it seems, only the unenlightened can pretend to be comfortable. If the neuroscientist were to restrict his investigations to his own clearly circumscribed field, and if he were to succeed in completely severing his understanding of that field from the understanding appropriate to the field of everyday human relationships, it would be less likely that he would see the activity of lots of neural processors as constitutive of thinking. It only becomes possible for him to view the neurobiological processes in the brain as constitutive of thinking if he widens the field of his inquiry, in order to postulate the type of unifying links demanded by the scientistic attitude. So it seems to me that it is precisely this attitude that accounts for the tendency to over-inflate the importance of the brain when it comes to explaining the individual's capacity to think, since it is precisely this attitude that gives rise to the overwhelming need to explain how the mental can be integrated into the physical dimension of human life.

3.2.The Consequence For Mental Autonomy

If this is correct, then the claim that mentalistic explanations must draw on the support of physicalistic explanations in carrying out their work can be said to be an explicit expression of the attitude of scientistic optimism. To suppose that mentalistic explanations cannot work autonomously is to suppose that the mental must be related to the physical in the manner suggested by the stand point of scientistic optimism. It is to suppose that mental states and events must be related to physical states and events by way of realisation or constitution, and that mental states and events can be legitimately cited in explanations of human action only because they derive their efficacy from the causal efficacy of these underlying physical states and events. Part of the process of coming to see that mentalistic explanations are autonomous in the deeper sense is the process of loosening the hold that the attitude of scientistic optimism has over our perception in these matters. To be in the position to see that mentalistic explanations can carry out their work without implicating the explanatory resources of the physical sciences is to be in the position to see that it is not necessary that thinking be related to the neurobiological functioning of the brain in the manner dictated by this attitude.

The reason why this is not necessary can be brought out by focusing on a way of expressing the dualism inherent in our concept of what it means to be a human being. This dualism can be understood as indicative of the fact that whilst human beings are part of the natural world,

in the straight-forward sense that human life has a basic biological constitution, human beings nonetheless learn to live in such a way that they come to have a life of their own, which serves to orientate them conceptually in the different situations in which they find themselves. Having a life to live in this sense is distinctive of what it means to be a human being, separating human beings from the rest of the living creatures who inhabit the natural world. The important point to note is that in living out their lives, human beings are sensitive to the different ways in which different people and different things present themselves to them, and in many cases, to the different ways in which they present themselves to other people. It is within this dimension of human life that we have to look to find the diversely over-lapping and loosely inter-locking contexts and situations that give sense to our various uses of mental concepts. As Dilman puts it:

> If we want to be clear about what it means to speak of the mind of a living thing, we need to turn our attention to those aspects of our life in which the reality of others for us finds expression, a life in which necessarily we ourselves are persons. (1996: 189).

In having a life to live, human beings are distinct from other living creatures in that they have their own personal psychology. But rather than look for ways of identifying aspects of that personal psychology with aspects of their neurobiological constitution, it seems to me that we have to recognise the implication of the fact that the externalisation of the mental is an inelimenable feature of the way of living distinctive of human beings. We have to recognise that thinking is embedded in the lives that human beings live in such a way that precludes its identification with what is going on in the brain. What we ought to say is that the focal point of our investigation is the human being, and that in having a life to live, the human being is no longer exclusively subject to the drives and forces of his biological nature, on a level with other non-linguistic but sentient creatures, but is rather *open*[4] to the world that is shaped and structured through the activities of other linguistic creatures who have a conceptually constituted orientation like himself. And this means that the connection between the mind and the brain is a connection that cannot be properly stated in terms integral to the non-reductivist's metaphysical framework, because the attempt to identify the mental with the physical is, I suggest, to misrepresent the distinction between mentalistic and physicalistic modes of explanation.

---

[4] The imagery of 'openness to the world' as a way of expressing that which is distinctive of human beings is developed in McDowell (1996), and also in Olafson (1995).

This distinction is one which is more properly expressed in terms of the fact that human beings have their own life to live in the world, and hence have their own personal psychology; which is to say that human beings, whilst being partly subject to the drives of their natural biological constitution, are at the same time partly subject to the different ways in which different people and different things *present* themselves to them in different situations. But the problem with the attempt to identify the mind with the brain is that it fails to acknowledge that the individual's own personal psychology is inseparable from his openness to the world. This is because it is part of having an openness to the world that factors in that world can figure constitutively in his thinking, and as the argument of the previous chapter suggested, this does not sit comfortably with the internalisation of the mental in the literally spatial sense required by that identification. So a more appropriate tact, in stating the connection between the mental and the physical, is to resist being drawn into the idea that we have to find a way of embedding the mental in the physical, and to consider instead how the brain serves its neurobiological function in sustaining human life, and hence how it serves its function as a causal enabling condition for the different ways in which human beings behave and respond as they live out their lives with each other on a day to day basis.

Thus, mentalistic explanations carry out their work within the lives that human beings live, whereas physicalistic explanations carry out their work within the neurobiological dimensions of human life. So the deep extent of mental autonomy can be said to consist not only in the fact that there is a logical gap between these dimensions of human life, and hence between the interests served by each of these modes of explanation, but also in the fact that the type of everyday relationships constitutive of the lives that human beings live are fundamental and basic. Which is to say that although mentalistic explanations presuppose that physicalistic explanations are already in place as part of the causal background, they do not carry out the work required of them through deriving their explanatory efficacy from physicalistic explanations in carrying out the work required of *them*. As I argued in the fourth chapter, if we assume that our everyday relationships necessarily begin from a non-everyday level, and subsequently work toward an everyday level through a process of interpretation, then our interest in others is necessarily an interest in causally explaining their behaviour by reference to the reasons they have for acting in certain ways. This is the crux of the matter. It is a consequence of accepting this assumption that the physicalist is forced to regard mentalistic and physicalistic explanations as converging on the common subject-matter of internal

behaviour-causing states and events. But although mentalistic explanations are certainly recognised as drawing on a system of irreducible concepts, the problem is that once the physicalist's metaphysical principles are firmly in place, dictating the extent to which the autonomous nature of mentalistic explanations can be respected, mentalistic explanations necessarily lose their deep autonomy.

4.Conclusion

Although I have been trying to exercise some amount of caution when it comes to understanding the role of the brain in explaining thinking, it has not been my intention to suggest that the brain is completely irrelevant. It has only been my intention to resist being caught up in the spirit of scientistic optimism, according to which the functioning of the brain must be given its prime status, as that with which thinking is to be identified. It seems to me that this is to over-inflate the role of the brain with regard to thinking, the only justification for which is the questionable methodological assumption that everything must be explained in physicalistic terms. The smooth and continued functioning of the brain is certainly relevant to an individual's having the capacity to think, in the sense that damage to certain parts of the brain often results in the sudden inability to exercise various cognitive and behavioural capacities with the same degree of spontaneity and integration as previously. But this is not to say that an individual suffering damage to certain parts of his brain would necessarily lose the capacity to think, since that would be to suppose that the smooth and continued functioning of the brain was all that mattered. Admittedly, it would be difficult to know what to say in such cases, since each individual would have to be considered on his own merits. Although there is an important respect in which having the capacity to think causally presupposes the continued functioning of the brain, this in itself does not warrant that conclusion that thinking goes on in the brain, or that the brain is our thinking thing; nor, for that matter, does it warrant the conclusion that mentalistic explanations must implicate physicalistic explanations in carrying out the work required of them.

Chapter 8: Animal Thinking

1.Introduction

It is an obvious merit of an account of thinking, or of an account of the mental in general, if it can be said to make sense of our tendency to treat some non-human animals as having the capacity to think, or as having a mind. What I have said so far, however, seems to leave me without an explanation of the fact that we do treat certain non-human animals as having the capacity to think. It seems as if I ought to be committed to denying that it makes sense to say that non-human animals can think, on the grounds that non-human animals cannot be said to have the capacity to be involved in human relationships. This presents me with a problem: in arguing that there is an intrinsic connection between what it means to have the capacity to think and what it means to have the capacity to be involved in human relationships, it might seem as if I have unwittingly denied that we can find a legitimate sense in which to attribute thoughts to non-human animals who, by definition alone, cannot be said to be involved in human relationships, despite the fact that we commonly explain why such animals are behaving in certain ways in mentalistic terms, by attributing to them the thoughts, intentions or desires which appear to rationalise their behaviour. So what I want to do in this chapter, to round off my argument, is argue that the intrinsic connection that I have been insisting on (between the capacity for thought and the capacity to be involved in human relationships) can be sustained, whilst doing justice to our irresistible inclination to treat certain animals (without being able to give a fixed list of which ones) as thinking animals.

2.Animal thinking: a further fragment of our already fragmented concept

2.1.McDowell 's Distinction: Living In A world And Living In An Environment

It might be useful to begin by considering an interesting distinction, which I alluded to in the previous chapter, between 'living in a world' and 'living in an environment'. This distinction is put to use by McDowell (1996: 114-9), in explaining how to make sense of an animal's sensitivity to certain features of its environment, without having to attribute to that animal the type of orientation in the world which it could only have through the possession of a language. Presumably, we can say that an animal that lacks a conceptually constituted

orientation in the world is an animal that lacks the full-fledged subjectivity which goes hand in hand with the capacity for thought. McDowell's idea is that we can only get to grips with an animal's mode of existence, including the form that its sensitivity to its environment takes, through getting to grips with the merely biological drives that shape its life. On this conception, an animal's life turns out to be no more than a succession of problems and opportunities, constituted as such by the biological forces that exert their control over its behaviour. Human beings, on the other hand, live in the world. Through being brought up into a particular way of living, which only becomes accessible through the acquisition of a language, human beings come to have the type of orientation in the world which non-human animals could never have. But this does not mean that animals are to be treated as senseless automata, for we need not credit animals with a conceptually constituted orientation in the world in crediting them with an alertness to their own environment.

McDowell's idea is that what sets human beings apart from animals is that human beings can come to live in a world through acquiring the capacity to conceptualise their situation, and this is a feat that animals can never achieve in so far as the patterns of their lives are not linguistically structured. I think that this is a useful way of bringing out a fundamental difference between the lives of human beings and the lives of animals. For the capacity to conceptualise a situation, and to adopt a rational response to the demands that it presents, seems to be constitutive of the lives of human beings in a way that it is not constitutive of the lives of animals. But by refusing to credit animals with this capacity, we also seem to be refusing to credit animals with the capacity to think. This brings us directly into conflict with our inclination to attribute thoughts and beliefs to animals in making sense of their behaviour, and it is not very clear how this conflict is to be resolved, particularly if we want to retain the distinction between living in a world and living in an environment.

2.2a.Malcolm's Distinction: Thinking And Having Thoughts

One way in which this conflict might be resolved is to appeal to the distinction between what it means to think, and what it means to have a thought. This distinction seems to be at the centre of Malcolm's (1977) efforts to make sense of the fact that we legitimately use the concept thinking with respect to animals, without implying that they can have thoughts before their minds. It is meant to capture the difference there is between merely thinking that such and such is the case, without consciously formulating the proposition that such and such is the

case, and actually having the thought before one's mind that such and such is the case, which does involve consciously formulating that proposition. Malcolm believes that by appealing to this distinction, which we commonly make in everyday life anyway, we can appreciate what is involved in attributing thoughts to animals, despite the fact that they lack the capacity to conceptualise their situation. Malcolm asks us to:

> Suppose our dog is chasing the neighbour's cat. The latter runs full tilt toward an oak tree, but suddenly swerves at the last moment and disappears up a nearby maple tree. The dog doesn't see this manoeuvre, and on arriving at the oak tree he rears up on his hind legs, paws the trunk as if trying to scale it, and barks excitedly into the branches above. We who observe the whole episode from a window say, "He thinks that the cat went up that oak tree." We say, "thinks" because he is barking up the wrong tree. (1977: 49-50).

Malcolm's idea is that we can be justified in attributing this thought to the dog, despite the fact that it lacks the capacity to formulate or entertain the proposition that the cat went up the oak tree, because there is a way of using the concept thinking which does not entail that the subject had a particular thought or that a particular thought occurred to the subject. This seems to be fair enough. We often do describe a dog's behaviour by saying that it thinks that such and such, or that it believes that such and such, but we wouldn't want to imply that the thought that such and such had occurred to the dog or that this thought went through its mind. Malcolm goes on to argue that we use the concept thinking in the same way with regard to people:

> suppose a friend of mine and I are engrossed in an exciting conversation. We are about to drive off in his car. While holding up his end of the conversation he fumbles in his pocket for the car keys. I, knowing that they are in the glove compartment, say to myself, "He thinks the keys are in his pocket." I do not imply that he said to himself, or thought to himself, "The keys are in my pocket." (1977: 49).

Malcolm seems to be arguing that since the man need not have the thought that his keys are in his pocket, just because it is correct to say that he thinks the keys are in his pocket, it follows that the dog need not have the thought that the cat was in the tree, just because it is correct to say that the dog thinks the cat is in the tree. But this is where the argument is in danger of breaking down. In so far as we use the concept thinking in both cases without meaning to imply that a thought had occurred to the man or the dog, there is a similarity in usage; but there is also an important dissimilarity in usage, which needs to be elucidated. The man who lost his keys is correctly described as thinking that his keys are in his pocket, and this description does not entail that he had the thought that his keys are in his pocket. So it looks as if we can say that the dog thinks the cat is up the tree, because this description does not

entail that the dog had the thought that the cat is up the tree. But this ignores the fact that the man who thinks that his keys are in his pocket has made a *mistake*, and that the use of the word 'think' on this occasion is tied to the possibility of his coming to realise that he has made a mistake. Perhaps the dog will notice the cat's tail hanging down from the branch of the maple tree, and immediately begin to bark up that tree; but even if we were to say that the dog made a mistake, as well we might, we would do so without expecting that the dog could ever come to realise that it had made this mistake. Unlike the friend who mistakenly thought that his car keys were in his pocket, the dog lacks the conceptual resources to be in the position to realise that it had been wrong. But unless we can attribute the possibility of coming to realise that a mistake has been made, it is not clear that this use of the word 'think' is the very same with respect to the man and the dog. This highlights a further point of difference between the use of the concept thinking with regard to the man and the dog. A logical precondition for saying that the man thinks his keys are in his pocket is that he could have had that thought before his mind. Although our description of the man does not entail that he *did* have the thought before his mind, it seems to entail that he *could* have had that thought before his mind. But this is not the case with the dog who is barking up the wrong tree, nor is it the case with the cat who is watching from the branch of the nearby maple tree. So if Malcolm's distinction is to be exploited as a way of resolving the conflict, which is underlined by McDowell's distinction between living in a world and living in an environment, it will have to developed somewhat in order to bring out significance of these differences. The development I have in mind will focus on the significance of the point that when we use the concept thinking with respect to certain animals, there is a sense in which we are *not* using it in the same way in which we use it with regard to people.

2.2b.Davidson 's Objection: Thinking, Believing And Being Surprised

Before that, however, it might be worthwhile considering an objection to this distinction in any form, which is raised by Davidson (1985). Davidson's objection is that, in saying that the dog thinks the cat ran up the oak tree, we are making an attribution of content which the dog's observed behaviour is not complex enough to support. The dog can only be said to think the cat ran up the oak tree if it can be said to have the appropriate background of beliefs to fix this content, namely, that trees are growing things, that they need soil and water, that they have leaves or needles, that they burn, and so on. Without the appropriately structured supporting background, which is inaccessible to non-language users, there would be no

justification for describing the dog's behaviour in this way. Davidson's objection thus leans heavily on the idea that the mental is holistic, and that the normative and rationalistic relations which govern this holistic network are of a type which can only be grasped by language users. The crux is that languageless creatures cannot be said to think because they cannot be said to have the appropriate background of beliefs, and they cannot be said to have that, because they cannot be said to understand what it would mean for its beliefs to be true or false, or justified or unjustified. Davidson tries to illustrate this logical requirement by considering the phenomenon of surprise. The idea is that unless a creature can be subject to surprises, in the sense that it can come to realise that what it previously believed was false, then it cannot be said to have any beliefs to begin with, and if it cannot be said to have any beliefs to begin with, then it cannot be said to have any thoughts either:

> Suppose I believe there is a coin in my pocket. I empty my pocket and find no coin. I am surprised. Clearly enough I could not be surprised (though I could be startled) if I did not have beliefs in the first place. And perhaps it is equally clear that having a belief, at least one of the sort I have taken for my example, entails the possibility of surprise. If I believe I have a coin in my pocket, something might happen that would change my mind. But surprise involves a further step. It is not enough that I first believe there is a coin in my pocket, and after emptying my pocket I no longer have this belief. Surprise requires that I be aware of a contrast between what I did believe and what I come to believe. Such awareness, however, is a belief about a belief (1985: 479).

The point is that beliefs are conceived as states which can be true or false. So unless a creature can be said to be aware of what it would mean for his beliefs to be true or false, that creature cannot be said to have any beliefs. This immediately excludes animals from what must turn out to be a well-defined domain of thinkers: only creatures that are equipped with a language can be said to have the capacity to think. It follows necessarily from this that it is incorrect to say that the dog thinks the cat ran up the tree. It might *look* as if there is nothing wrong with saying this, given that it seems to be justified by the dog's behaviour on this particular occasion; but when we take into account the fact that the identity of any one thought is dependent on its location within the logical network of content-determining beliefs, it becomes difficult to make sense of our attributions of thoughts about the cat and the tree to the dog.

Davidson's objection certainly seems to put pressure on Maclolm's distinction, and I have to agree that the problem concerning the determination of content does seem to pose a *prima facie* challenge to the claim that the dog can think the cat ran up the tree. But it seems to me that there is something not quite correct with the idea that we can justify this attribution of content to the dog only if we presuppose that it has access to the distinction between truth and

falsehood that language users have. Perhaps Malcolm invites this objection, however, in pointing out that we use the concept thinking in the *same way* with regard to animals and people, but I want to avoid it by suggesting that we use the concept thinking with regard to animals in a slightly *different way*.

2.3.Our Relationships With Animals: The Distinction Developed

According to Gaita (1992: 237), there is a tendency to suppose that the right method in investigating animal thinking is to distance ourselves from our subject so that our emotions and affections are prevented from interfering with our sense of what is 'objectively the case'. This tendency manifests itself in the need to find a language for describing animals which is only contingently vulnerable to various aspects of human nature, to our passions, our fantasies and our failings of character. This seems to highlight a submerged strand in Davidson's position, where the question of whether animals can be said to think is to be settled in an *a priori* manner. The central issue for Davidson is whether animals can grasp the distinction between objective truth and falsehood, and whether their behavioural repertoire is sufficiently complex to support this distinction. His claim is that only linguistic behaviour will suffice, and that languageless creatures therefore lack the capacity to think.

But Gaita points out, with regard to human beings first of all, that what gives mental concepts their sense cannot be separated from our emotional and affective natures, and that it is precisely this side of human life which endows us with those modes of understanding which are appropriate to 'seeing the reality' of another person. He goes on to suggest that there is no reason to suppose that things must be radically different in relation to animals, in the sense that it is a mistake to suppose that we can determine 'objectively' whether or not animals can think by adopting a criterion which forces us to disengage from our emotional and affective natures. In Gaita's view, we can only get to grips with what is objectively the case *in these issues* through the recognition that certain forms of emotion and affection are themselves modes of understanding, and this means that *what is objectively there* to be understood cannot be characterised independently of the fact that our only access to this type of reality is through precisely these emotions and affections. Or in other words, we cannot get to grips with animal thinking if we ignore the importance of our emotional and affective orientation toward animals in settling the issue, since it is precisely that orientation that gives sense to our mental concepts.

This suggests a way of developing Malcolm's distinction, and hence resolving the conflict with which this chapter started: on the one hand, we are inclined to attribute thoughts and beliefs to certain animals in explaining their behaviour; but on the other hand, it seems we ought to deny that animals can think, given that they lack the capacity to be involved in human relationships, and that they lack the conceptual capacities which are required if they are to be credited with anything like the orientation in the word that comes hand in hand with having the capacity for thought. There need be no irresolvable conflict here if we can come to appreciate the connection between the way we relate to animals and the fact that we sometimes explain their behaviour in mentalistic terms. For as Gaita notes, our relationships with some animals can involve a great deal of emotional and affective interaction, and can verge on matters of moral importance: we love and care for animals, we take care of them and train them; they seek our love and affection, they engage our attention, pity and respect; they play with us, and sometimes comfort us by being there. And it seems to me that in relating to animals in these ways, we are involving them in certain aspects of our lives, and we are involving ourselves in certain aspects of their lives, and we therefore become significant parts of their environment *in making them* significant parts of our world.

To the young child, for example, the family dog might simply be another play mate, and the source of joy and amusement; but to the older brother, it might rather be the source of great annoyance and irritation. To the blind person, the guide dog is not simply a play mate, nor is it simply a pet to be fed and watered on a daily basis; it is rather indispensable to his carrying on his life in a near normal manner, almost as an extension of his own body. To the elderly man, perhaps house-bound through illness, his dog might be the only companion he has when his nurse or carer goes home at night. Certain animals thus develop their own personalities, and some of them become irreplaceable, through the different ways in which they become involved in our lives. The child might not be content with a replacement play mate, although his brother might be happy to have a quieter pet. The blind person would probably need a lengthy period of readjustment before he was relaxed and comfortable with his new guide dog, and even then he would probably spend a great deal of time thinking affectionately about his previous dog. The elderly man would feel the loss of his life companion in a deep way, and would lose a part of his life that was of great importance to him.

It seems to me that attributing thoughts to animals is part of the way we in which relate to them, when we play with them, feed them and train them, when we seek their comfort and

their security.[1] Given that animals are capable of relating to each other in certain ways, and that some are capable of soliciting responses from *us* which are similar to the responses which other people solicit from us in similar situations, we cannot help attributing thoughts to animals. Some of us find it *natural* to talk directly to certain animals, those which have a human-like face, to look into their eyes as we look into the eyes of another person, or to tell stories and talk about them to other people, using an array of mental concepts, including thinking, intending, believing, desiring, and so on. But it seems to me, and this is the important point, that the way in which we use the concept thinking in relating to animals is not quite the same way in which it is used in relating to other human beings. The way we use the concept thinking in relating to animals does not carry the same logical baggage that is carried by the way we use it in our relationships with other people.

It seems to me that this is the point that Malcolm should have made when he made the distinction between saying that animals can think that such and such is the case, without having the thought that such and such is the case before their minds. Given that our relationships with animals do not have the same types of complications and demands as our relationships with other people, there is no need for the use of the concept thinking in relating to animals to carry the same amount of logical baggage. We do not expect Malcolm's dog to be aware of having made a mistake when it noticed the cat's tail hanging down from the maple tree, nor do we expect it to be aware of having been duped for the third time in succession by the same mischievous cat; nor do we expect it to be aware of the fact that the oak tree is the oldest tree in sight, or that the maple tree is the one the cat disappeared into the time before and the time before that. The different ways in which the concept thinking is used with regard to animals and people is registered in the different ways in which we relate to animals and people. Many of the complications and demands which are constitutive of our relationships with other people are not constitutive of our relationships with animals, even those animals which are closest to us. But this means that we can legitimately use the concept thinking in relating to animals without logically implying that they are capable of having thoughts before their minds, and hence without having to agree that we are using a mode of explanation that goes beyond what the situation and their behaviour merits.

---

[1] In making this point I do not mean to deny that it makes sense to talk of wild animals as thinking. Although we are not involved in relationships with them, they are not so radically different from trained domestic animals in their behaviour and appearance that we would find it unnatural to use mental concepts in understanding their behaviour. We study wild animals from a distance as interested spectators, when we stand back and watch them at *play* with other animals, as they *mimick* each other's bodily movements, or as they *follow the lead* of the dominant one, as they engage in patterns of behaviour and relationships natural to both humans and animals alike, and so on, and so forth.

So we might understand our use of the concept thinking with regard to animals if we see it as determined by the form of these *particular* relationships, and this means that we have to acknowledge that we are using it in a slightly different sense with regard to animals from the senses in which we use it with regard to other people. So rather than say that we cannot use the concept thinking with regard to animals because they cannot be involved in human relationships, we might say that in using the concept thinking with regard to animals, we are using it in a manner appropriate to the forms of their relationships, both with human beings and with other animals. The different uses we make of the concept thinking with regard to animals and people is therefore registered in the different forms that our relationships take, and this should be seen as highlighting a further fragment of the concept thinking, in addition to its already fragmented uses with regard to other people. The point of recognising this particular use of the concept thinking is that it allows us to see how mentalistic explanations can be legitimately used with regard to certain languageless creatures, even though such creatures do not live in a world, in the sense of being responsive to those features of their situations that present them with reasons for acting in certain ways, or feeling certain things. So we can maintain the distinction between living in a world and living in an environment, and we can maintain the intrinsic connection between having the capacity to think and having the capacity to be involved in human relationships, and yet we can also continue to explain animal behaviour in mentalistic terms, without having to admit that we are not quite correct to do so.[2]

---

[2] Computers and machines seem to present a further problematic case for my thesis; the obvious objection is to say that as technology advances further and further, it will eventually be beyond question to say that computers and machines can think. But in an interesting discussion, Dreyfus (1994) raises a number of objections to the idea that computers and machines can think, most of which stem from the central claim that thinking presupposes having a background of common-sense knowledge which cannot be encoded into the format required to drive computer and machine programmes. He makes much of the fact that thinking requires having the ability to zero-in on salient features of our situation and to ignore features which might have been salient given a different set of circumstances, to make relevant generalisations between situations, to learn from mistakes, and so on, all of which presuppose that we are in actual fact 'in a situation', that we have the needs, interests and concerns, in terms of which features of our situation *are* salient or relevant, and which cannot be had by *disembodied* computers and machines. This seems correct, but it also seems correct to say that if the computers and machines (and this goes for aliens too) were embodied, if they had the skills and abilities that disembodied ones lack, *if they could interact with the rest of us in a near enough ordinary manner, then* it would no longer be clear-cut whether they could be said to think or not. For as Hanfling (1991) rightly points out, if such machines had an array of personal and moral qualities, if they could make various types of demands on us, and hence if they were 'persons' in this respect, then it would only be a prejudice (*artifactism*, to be precise) to refuse to say that they could think. Indeed, if we regard the word 'think' as part of the way in which we relate to each other, as I have been arguing, then it might turn out to be an effort to resist taking it for granted that Hanfling's machines could think.

Chapter 9: Conclusion

The chief aim of this thesis has been to argue that the autonomous nature of mentalistic explanation imposes a strong constraint on what counts as a satisfactory statement of the relation between the mental and the physical. It has been argued that there is a deeper extent to the autonomous nature of mentalistic explanation than can be appreciated within the metaphysical framework of non-reductive physicalism. Within that framework, the autonomous nature of mentalistic explanation turns out to be a matter of the irreducibility of mental concepts to physical concepts. But I have claimed that this is not sufficient. In addition to the requirement that mental concepts are irreducible to physical concepts, there is a stronger requirement which cannot be met within this metaphysical framework. The stronger requirement is that a satisfactory statement of the relation between the mental and the physical must be sensitive to the fact that mentalistic explanations are autonomous in the much deeper sense that they can carry out the work required of them without having to derive explanatory support from physicalistic explanations in carrying out the work required of *them*.

I argued in the second chapter that the basic reason why the deeper extent of mental autonomy has not been appreciated is that the metaphysical principles integral to the non-reductivist's framework yield a statement of the relation between the mental and the physical which is inconsistent with it. The consequence of adopting these principles is that the mental has to be seen to be embedded in the physical structure of the world, and mentalistic explanations must therefore converge on the very same subject matter as physicalistic explanations, in the sense that both explain the occurrence of the same events, albeit using different modes of description. Once the mental is related to the physical in the manner demanded by the metaphysical principles of physicalism, the most that can be acknowledged is that the modes of description and explanation which converge on the same events in fact *diverge* in their respective methodologies. But the basicness and completeness of the explanatory resources of the physical sciences make it impossible for mentalistic explanations to work successfully on their own, since whatever work is carried out by mentalistic explanations will have to be supported by the underlying physicalistic explanations of these same events. So whereas the metaphysical framework of non-reductive physicalism seems suitably structured to recognise the requirement that mentalistic

explanations are irreducible to physicalistic explanations, its principles impose an ordering onto reality that prevent it from recognising the fact that mentalistic explanations can carry out their work without having to derive explanatory support from physicalistic explanations.

In the fifth chapter, where I discussed the nature of rationalising explanations, I argued that matters might actually be worse than this. I argued that the metaphysical framework of non-reductive physicalism comes dangerously close to undermining mental autonomy altogether, since doubts have arisen over the causal efficacy of mental properties in the physical world. The problem is that the non-reductivist can secure mental autonomy only in so far as the rationalising mode of explanation is guaranteed to be irreducible to the physicalistic mode of explanation, but in order to combine this aspect of mental autonomy with the principles of physicalism, he must construe rationalising explanations as a species of causal explanation, where the important causal processes are implemented at the physical level. But if it is correct to argue that mental properties can be regarded as real and autonomous features of the world only in so far as they inherit their causal efficacy from the underlying physical properties which constitute their realisation base, then it follows that mental properties have no independent causal role to play in the physical world. And if this is the case, then rationalising explanations cannot be construed as a species of causal explanation in the desired sense, without risking the loss of their autonomy to the physicalistic mode of explanation altogether. There seems to be no genuine explanatory work left for the rationalising mode of explanation to carry out once the physicalistic mode of explanation has carried out its work, and the mental thereby threatens to be nothing more than an epiphenomenal feature of the physical world.

My position is that if we take the contexts of our *everyday relationships* as the focal point of our investigation into the nature of our explanatory practices, then we can begin to meet the requirements imposed on us by the autonomous nature of mentalistic explanations. We can do this because we will be able to determine whether mentalistic explanations are successful by determining whether they satisfy the understanding sought by each individual in his or her own relationship. What counts as a successful mentalistic explanation will be judged relative to the contexts of our involvement in relationships with each other, and hence the question of the success of mentalistic explanations in carrying out the work required of them will be independent of the question of the success of physicalistic explanations in carrying out the work required of them. However, I pointed out that it was not sufficient for my argument to

claim that mentalistic and physicalistic explanations have different criteria for success, since the mere fact that mentalistic explanations carry out their work according to different standards does not establish that mentalistic explanations are autonomous in the deeper sense. It could still be the case that mentalistic explanations have to derive explanatory support from physicalistic explanations, even though their success in carrying out their work is assessed according to different criteria.

For this reason, I pointed out that it would be necessary to demonstrate that mentalistic explanations can indeed work separately from physicalistic explanations, and for that purpose I developed a non-causal approach to rationalising explanations. The non-causal approach was built on the intrinsic connection I drew between the nature of thinking and the nature of human relationships. The main point was that rationalising explanations explain our actions in terms of our responsiveness to the demands of the situations in which we find ourselves, which they can do without implicating the explanatory resources of the physical sciences. It is only if we must assume that the notion of causality, in the particular sense required by the physicalist, must be built into the notion of what it means to act for a reason, that we are forced to say that mentalistic explanations cannot carry out their work independently of physicalistic explanations.

The upshot of all of this is that once we reject the metaphysical framework of non-reductive physicalism, and take our successful explanatory practices as our starting point, we can meet both requirements for recognising mental autonomy: first of all, mentalistic explanations are irreducible to physicalistic explanations because the success of mentalistic explanations in carrying out their work is assessed relative to their ability to satisfy the understanding we seek in our everyday relationships with each other, which is distinct from the understanding sought by neurobiologists in studying the functioning of the brain and the human nervous system; second, mentalistic explanations can carry out their work successfully without deriving explanatory support from physicalistic explanations in carrying out their work, since mentalistic explanations do not explain the same thing as physicalistic explanations, namely, the causation of behaviour by internal physical events and processes.

The central claim of this thesis is that a satisfactory statement of the relation between the mental and the physical is one which is constrained by the autonomous nature of mentalistic explanation in the deep sense. Therefore, I suggest that in stating this relation, we should take

our cue from the considerations which have put us into position to meet the two requirements outlined above. What this means is that we should avoid trying to find a means of articulating the relation between the mental and the physical that would enable us to reconcile the embeddedness of the mental in the physical structure of the world with the irreducibility of mental concepts to physical concepts. We should rather try to conceive this relation in such a manner that is consistent with the distinction between mentalistic and physicalistic explanations as they are they are ordinarily employed in our everyday explanatory practices. This distinction can be respected if we think of it as expressing the dualism implicit in the concept of what it means to be a human being: that human beings are subject to physicalistic explanation is to say that human beings are creatures whose life has a natural biological dimension; that human beings are subject to mentalistic explanation is to say that human beings have a life to live in the world with other people.

The relation between the mental and the physical might therefore be understood in terms of the fact that the functioning of the brain and the rest of the nervous system serves as a causal enabling condition for human beings to live their lives in the world with each other. As such, it is not correct to say that the mental is embedded in the physical structure of the world; rather is it embedded in the world in which human beings live their lives.[1] This world is constituted as such through the involvement of human beings in various types of relationships with each other. For a human being to have a mind is for him to have a life to live in the world. For a human being to have a life to live in the world is for him to have a conceptually constituted orientation in the different situations in which he finds himself, which has been developed and shaped into a stable and lasting outlook through the long and complicated process in which he learned to relate to others. This process of development is not a causal process, characteristic of the physical dimension of human life; rather is it a normatively constrained process of concept acquisition, characteristic of the dimension of human life in which we are involved in various types of relationships with each other. It is the latter dimension of human life which provides the contexts in which the mental can be said to be embedded, and the physical dimension of human life can be said to sustain our life in the world by making various features of it causally possible.

---

[1] As Wittgenstein (1980b: § 16) neatly puts it: "In this case I have used the term "embedded", have said that hope, belief, etc., were embedded in human life, in all the situations and reactions which constitute human life."

In order to articulate the relation between the mental and the physical more clearly than this, it would be necessary to engage in an empirical inquiry into the different ways in which the brain serves as an enabling condition that makes our lives in the world causally possible. But however this inquiry should turn out, I suggest that it ought to be conducted with a sensitivity to the constraint presented by the autonomous nature of mentalistic explanation in the deeper sense. So rather than focusing on a way of incorporating the mental into the physical structure of the world by means of unifying principles which purport to justify the claim that mental states and events just are physical states and events, the focus ought to be restricted to a way of explaining how the functioning of the brain and the nervous system plays a causal role in sustaining the individual's everyday activities. To take the former approach is to force a structure onto the world whose upshot is to interfere with the way in which our mentalistic explanations are actually used. To take the latter approach is to recognise the soundness of the methodological point that we ought to take our inquiry no further than is warranted by the nature of our everyday explanatory practices. And this, I have been arguing, is to recognise the strong constraint imposed by the deep extent of the autonomous nature of mentalistic explanation.

Perhaps such an inquiry would reveal a limitation in my position, however, since I have not taken into consideration the fact that the treatment of certain mental disorders involves the use of drugs to control the release of specific chemicals in the brain, for instance, which suggests that there are points at which mentalistic explanations of these sorts are directly supported by physicalistic explanations. We can successfully explain why some individuals are mentally unstable by reference to their neurobiological condition, and we can temporarily exert some control over their mental condition by effecting the relevant changes in their neurobiological condition. We can also explain why an individual is suddenly afraid of the candlestick on the mantelpiece, why he thinks that he has the ability to fly, or to make himself invisible at will, by reference to the chemical changes in his brain effected by the intake of hallucinogenic drugs. What this suggests is that there is a difference between the type of mentalistic explanations employed in our everyday relationships with each other, and the type of mentalistic explanations employed in the treatment of mental disorders, or in the context of abnormal behaviour induced by interfering with the levels of chemicals released in the brain. So the constraint presented by the autonomous nature of mentalistic explanation in the deep sense will have to be relaxed somewhat in these circumstances; but I am not sure that I

could provide a means of clearly distinguishing between these cases, and the cases in which I have argued that the constraint should be more strictly adhered to.[2]

What I have also claimed in this thesis is that there is an intrinsic connection between the nature of the mental and the nature of our involvement in various types of human relationships, which is to say that the meaning of our mental concepts is inseparable from the circumstances and situations in which they are used as part of the way in which we relate to each other. To have a grasp of the meaning of our mental concepts is to know how to relate in endless ways to different people; it is to know how to cope with the demands that other people present to us throughout the course of our lives. There is no gap between what it means to know the meaning of mental concepts and what it means to know how to relate to other people. What we mean when we talk about the mental is thus intrinsically connected to what we mean when we talk about our involvement in relationships with other people, which is to say that our talk about the mental is talk about the way in which we live out our lives in the company of others. It might even be said, therefore, that our mental concepts *are* the ways in which we relate to each other, when we are moved by another person's expression of emotion, when we are touched by their declaration of affection, when we state our concerns and intentions, or quite simply when we show an interest in certain aspects of each other's lives.[3]

Mentalistic explanations are thus geared toward understanding others as responding to the demands of the situations in which they find themselves; they are geared toward understanding others as living out their lives in the world which has been shaped and structured through their involvement in various types of human relationships. Mentalistic explanations are subject to the normative and rationalistic constraints generated within the contexts of human relationships, and successful mentalistic explanations are those which satisfy the understanding sought by the individuals concerned within the context of their own particular relationship. But there is no need to regard mentalistic explanations as a species of

---

[2] Perhaps the distinction might be drawn in terms of the distinction between everyday and non-everyday relationships. It could be said that the concepts used in discussing the effects of using certain types of drugs in the treatment of mental disorders are explanatory concepts whose meanings are fixed by their usage in these specific medical contexts. Which is to say that although the uses of these concepts in explaining behaviour are directly tied to the individual's neurobiological condition, this is justified within my position by the fact that they carry out their explanatory work within the contexts of non-everyday relationships.

[3] Or, as Hertzberg (1983: 106) puts it: "the way human expressions move and affect us...in a sense, is what *constitutes* our mental concepts."

causal explanation, since that move is motivated by the idea that it is the only way to guarantee the fact that what an individual thinks is explanatory of what he does. But this is not what we should say. What we should say is this: the fact that what an individual thinks is explanatory of what he does is guaranteed by the fact that *this is how* we actually explain each other's behaviour.

This suggests that the concept of thought might be regarded as an explanatory concept, which finds application within the dimension of human life that is constituted by our involvement in various types of human relationships. The fact that what an individual thinks is explanatory of what he does is a grammatical fact concerning *what we mean* when we talk about thinking and acting, and what we mean when we talk about thinking and acting is exhaustible within the contexts in which we make use of these concepts in relating to each other. To suppose that there has to be certain metaphysical connections between the mental and the physical, in order that what an individual thinks can be said to be explanatory of what he does, is to overlook the point that we are concerned with a grammatical explanation of these concepts. The metaphysical framework of non-reductive physicalism treats this grammatical fact as standing in need of further justification, and in implementing that framework our explanatory practices suffer the distortion that appears to prevent mentalistic explanations from being completely adequate to their own tasks. But if we regard the fact that thinking explains acting as grammatical, then we need not attempt to ground it metaphysically. It is rather an explanation of what we generally mean when we talk about the uses of the concepts thinking and acting in everyday situations, as when we say, for example, that the trecker is checking his map because he thinks he might have taken a wrong turning somewhere, or that the riot police have called for military reinforcements because they think the crowds are getting out of control.

The intrinsic connection between the nature of the mental and the nature of human relationships also highlights an essential moral dimension to our talk about the mental which might be further developed along the following lines. To be responsive to the demands that other people present to us in our relationships with them is, among other things, a moral achievement. Through the proper cultivation of our basic attitudes, through the experiences we have at different times in our lives, we learn to see others as making various demands on us, as constraining our natural inclinations and tendencies, as obligating us to treat them in certain ways. We develop an outlook on our lives through the guidance we receive from our

parents, and later from our teachers and friends, which not only serves to define us as the individuals we are, but which at the same time sustains and shapes our relationships with others. We are at the same time actively involved in shaping that definition ourselves, when we undertake to improve or correct aspects of our lives we are not satisfied with, and sometimes more passively, through the subtle changes to our moods and attitudes which may result when we are drawn into revealing different aspects of our lives to different people. The ways in which other people respond to us contributes to our conception of ourselves, and the ways in which we conceive ourselves contributes to the sense of worth and value which we subsequently bring to our relationships with other people. The moral aspect of our mental concepts is brought out by their uses within the contexts of our relationships, which we learn to use correctly in the course of learning to relate to others: we regard an individual in pain as someone toward whom pity is an appropriate attitude to have; we regard an individual in distress as someone toward whom comfort is an appropriate attitude to have.

Part of what is involved in understanding the meaning of our mental concepts is understanding the fact that other people make demands on us, which we are responsive to through having developed the attitudes toward others which sustain our moral vision in these cases. Our moral vision is not a purely cognitive one, however, which provides us with an understanding of how we ought to respond to others by providing us with a normatively constrained interpretation of what happens to be the case. Rather is our moral vision inseparable from our emotional and affective natures, which provides us with the type of orientation toward others that makes such a task of interpretation seem superfluous. Leaving room to accommodate exceptions, it can be said that our understanding of what happens to be the case is already our understanding of how we ought to respond to others, because the concepts drawn on in specifying the content of our moral judgements are inseparable from the ways in which we actually relate to each other in the normal course of things. Sometimes we do not respond as we know we ought to, and occasionally our judgements are distorted or clouded by various factors; but this is not to detract from the point that the factual judgement *that the individual is in pain* draws on concepts whose normative implications are already in place, presenting us with a reason to offer him comfort, care, pity and so on.[4]

---

[4] According to Luntley (1991: 179-181), the normative implications grasped in circumstances like this one are in fact constitutive of our experiences. This is an interesting point. It supports the claim that in perceiving that an individual is in pain, we are immediately presented with a reason to respond to that individual in a particular manner. Our perception of the fact that he is in pain is at the same time our perception of the fact that a certain course of action is appropriate in these circumstances.

Finally, in claiming that there is an intrinsic connection between the nature of the mental and the nature of human relationships, I have used the notion of *presence* to express the way in which features of the world in which we live can figure in our thinking as giving us reasons for acting in certain ways and feeling certain things. What this thesis has not touched on, which is certainly relevant to the plausibility of the latter claim, is the question of whether this same notion can be used to develop a non-causal account of what it means to perceive something. To suppose that our ordinary concept of perception must somehow express the fact that a causal transaction is taking place between object and perceiver is indeed tempting, given that we would not be able to see or hear something if it did not bring about a neurophysiological change in our bodies. The central issue to be addressed would be whether the causal transaction between the object and the perceiver at the level of neurophysiological change has to be imported into our ordinary understanding of what we mean when we say that an individual perceives something.

One way in which the causal theory has been motivated, leaning more on epistemological than physiological concerns, has been to appeal to the idea that perception is a way of informing ourselves about the world of independently existing things.[5] The crux of the argument is this: if we think of perception as a way of informing ourselves about the way things are in the world independently of the way we take things to be, then we are forced to admit that the notion of causality must be built into the concept of what it means to perceive something. We have to assume the causal dependence of our perceptual experiences on the independently existing things we take them to be of, otherwise our experiences could not be said to inform us about the way the independently existing world happens to be. And if we do assume this causal dependence as necessary to explaining the reliability of the information conveyed by our perceptual experiences, then it follows that the notion of causality must be built into our ordinary conception of what it means to perceive something.

What is problematic about the causal theory of perception is that it assumes that there is a merely *extrinsic* connection between our perceptual experiences and the objects of which they are experiences. Or in other words, it assumes that the experience of seeing the cat jump onto the table, for instance, is extrinsically connected to the fact that the cat is jumping onto the table, and that the experience only has the content it has in virtue of the way it has been caused in the perceiver. This account of perception does not mesh neatly with the account of

---

[5] This particular route to the causal theory of perception is outlined by Strawson (1979).

thinking I have developed, since it posits an extrinsic connection between our perceptual experiences and the objects we perceive which would cohere rather neatly with the basic ideas expressed in the physicalist's conception of thought. The internal event would be the perceptual experience, and the external factor would be the object perceived. But if we can develop an account of perception, which appeals to the notion of presence as a way of expressing the *intrinsic* connection between our perceptual experiences and the objects we perceive, then we will be able to think of perceiving as a unitary act in which the independently existing object figures as a constituent part.[6] This would block an important escape route for the physicalist, since it would block the possibility of separating the perceptual experience from the object perceived in the required sense. But more importantly, it would suggest a means of extending some of the ideas developed in my approach to thinking and acting to cover our perceptual capacities. The larger aim of extending the results of this thesis to cover our perceptual capacities, which I would consider satisfactory, would be to provide a unified account of thought, perception and action, which would not only allow us to think of human beings as subject to both mentalistic and physicalistic explanations, but which would allow us to do this whilst respecting the deeper extent of the autonomous nature of mentalistic explanations.

---

[6] For an example of how perception might be understood in terms of presence, see Olafson (1995: 46-86); but see also the disjunctive conception of vision as outlined in Snowdon (1980-1), and McDowell (1988). And for a more recent debate over the causal and non-causal theory of perception, see Hyman (1992) and Child (1992).

Bibliography

First date given is that of edition or publication cited in text. Dates given in parentheses are those of first edition or publication if different.

Anscombe, G.E.M., 1968 (1956-7): Intention, in *The Philosophy of Action*, White, A., ed., 144-152. Oxford University Press.

Baker, L.R, 1993: Metaphysics and Mental Causation, in *Mental Causation*, Heil, J., and Mele, A., eds., 75-95. Oxford University Press.

Baker, L.R., 1995: *Explaining Attitudes: A Practical Approach To The Mind.* Cambridge University Press.

Baker, G.P. and Hacker, P.M.S., 1992 (1985): *Wittgenstein: Rules, Grammar and Necessity, Vol.2 of an Analytical Commentary on the Philosophical Investigations*. Basil Blackwell.

Brueckner, A., 1990: Scepticism about Knowledge of Content, *Mind*, Vol. 99, 395: 447-451.

Burge, T., 1979: Individualism and the Mental, in *Midwest Studies in Philosophy*, Vol. IV, French, P.A., Uehling, T.E. Jr., and Wettsein, H.K., eds., 73-121.

Burge, T., 1994 (1988): Individualism and Self-Knowledge, in *Self-Knowledge*, Cassam, Q., ed., 65-79. Oxford University Press.

Canfield, J.V., 1994: The Phenomena of Thinking, in *Wittgenstein and Contemporary Philosophy*, Teghrarian, S., ed., 109-130. Bristol, Thoemmes Press.

Child, W., 1992: Vision and Experience: The Causal Theory and the Disjunctive Conception, *The Philosophical Quarterly*, Vol. 42, No. 168: 297-316.

Child, W., 1994: *Causality, Interpretation, and the Mind*. Oxford University Press.

Cockburn, D., 1985: The Mind, the Brain and the Face, *Philosophy*, 60: 477-493.

Cockburn, D., 1990: *Other Human Beings*. MacMillan.

Davidson, D., 1980a (1970): Mental Events, in his *Essays on Actions & Events*, 207-227. Oxford University Press.

Davidson, D., 1980b (1963): Actions, Reasons, and Causes, in his *Essays on Actions & Events*, 3-19. Oxford University Press.

Davidson, D., 1984a (1975): Thought and Talk, in his *Inquiries into Truth & Interpretation*, 155-170. Oxford University Press.

Davidson, D., 1984b (1973): Radical Interpretation, in his *Inquiries into Truth & Interpretation*, 125-139. Oxford University Press.

Davidson, D., 1985 (1982): Rational Animals, in *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, LePore, E., and McLaughlin, B., eds., 473-80. Basil Blackwell.

Davidson, D., 1994 (1987): Knowing One's Own Mind, in *Self-Knowledge*, Cassam, Q., ed., 43-64. Oxford University Press.

Dilman, I., 1987: *Love and Human Separateness*. Basil Blackwell.

Dilman, I., 1990: Self-Knowledge and the Reality of Good and Evil, in *Value & Understanding*, Gaita, R., ed., 194-215. Routledge.

Dilman, I., 1996: Existence and Theory: Quine's Conception of Reality, in *Wittgenstein & Quine*, Arrington, R.L., and Glock, H.J., eds., 173-195. Routledge.

Dreyfus, H., 1994 (1972): *What Computers Still Can't Do*. M.I.T. Press.

Enç, B., 1995: Nonreducible Supervenient Causation, in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 169-186. Cambridge University Press.

Flanagan, O., 1992: *Consciousness Reconsidered*. M.I.T. Press.

Fodor, J., 1992 (1980): Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology, in *The Philosophy of Science*, Boyd, R., Gasper, P., and Trout, J.D., eds., 651-669. M.I.T. Press.

Fodor, J., 1981 (1974): Special Sciences, in his *Representations: Philosophical Essays on the Foundations of Cognitive Science*, 127-145. Brighton, Harvester Press.

Fodor, J., 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* M.I.T. Press.

Gaita, R., 1992: Animal Thoughts, *Philosophical Investigations*, 15, 3: 227-244.

Georgalis, N., 1990: No Access for the Externalist: Discussion of Heil's 'Privileged Access', *Mind*, Vol. 99, 393: 101-108.

Glock, H.J., 1993: The Indispensability of Translation in Quine and Davidson, *The Philosophical Quarterly*, Vol. 1, 43, No. 171: 194-209.

Glock, H.J., and Preston, J., 1995: Externalism and First Person Authority, *The Monist*, Vol. 78, No. 4: 515-533.

Goldfarb, W., 1992: Wittgenstein on Understanding, in *Midwest Studies in Philosophy*, Vol. XVII, French, P.A., Uehling, T.E. Jr., and Wettsein, H.K., eds., 109-122.

Grimes, T., 1995: The Tweedledum and Tweedledee of Supervenience, in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 110-123. Cambridge University Press.

Hacker, P.M.S., 1993 (1990): *Wittgenstein: Meaning and Mind, Volume 3 of an Analytical Commentary on the Philosophical Investigations, Part 1: Essays*. Basil Blackwell.

Hacker, P.M.S., 1996: *Wittgenstein: Mind and Will, Volume 4 of an Analytical Commentary on the Philosophical Investigations*. Basil Blackwell.

Hampshire, S., 1976 (1960): Feeling and Expression, in *The Philosophy of Mind*, Glover, J., ed., 73-83. Oxford University Press.

Hanfling, O., 1991: Machines as Persons? in *Human Beings: Royal Institute of Philosophy*, 29: 25-34.

Hanfling, O., 1993: '"Thinking", a widely ramified concept', *Philosophical Investigations*, 16, 2: 101-115.

Heidegger, M., 1967 (1962): *Being and Time*, Macquarrie and Robinson trans. Blackwell.

Hellman, G.P., and Thompson, F.W., 1975: Physicalism: Ontology, Determination, and Reduction, *The Journal of Philosophy*, 72: 551-564.

Hertzberg, L., 1983: The Indeterminacy of the Mental, *Proceedings of the Aristotelian Society*, Supplementary Volume, LVII: 91- 109.

Honderich, T., 1982: The Argument for Anomalous Monism, *Analysis*, 42: 59-64.

Hornsby, J., 1980-1: Which Physical Events are Mental Events? *Proceedings of the Aristotelian Society*, Vol. LXXXI: 73-92.

Hunter, J., 1987: Some Thinking about Thinking, *Philosophical Investigations*, 10, 2: 118-133.

Hunter, J., 1990: *Words as Instruments*. Edinburgh University Press.

Hyman, J., 1992: The Causal Theory of Perception, *The Philosophical Quarterly*, Vol 42, No. 168: 277- 296.

Jackson, F., and Pettit, P., 1993: Some Content is Narrow, in *Mental Causation*, Heil, J., and Mele, A., eds., 259-282. Oxford University Press.

Kim, J., 1993a (1989): The Myth of Non-Reductive Materialism, in his *Supervenience and Mind*, 265-284. Cambridge University Press.

Kim, J., 1993b (1990): Supervenience as a Philosophical Concept, in his *Supervenience and Mind*, 131-160. Cambridge University Press.

Kim, J., 1993c: The Non-Reductivist's Troubles With Mental Causation, in his *Supervenience and Mind*, 336-357. Cambridge University Press.

Kenny, A.J.P., 1985 (1971): The Homunculus Fallacy, in his *The Legacy of Wittgenstein*, 125-136. Basil Blackwell.

Kripke, S., 1982: *Wittgenstein on Rules and Private Language*. Basil Blackwell.

Luntley, M., 1991: The Transcendental Grounds of Meaning and the Place of Silence, in *Meaning Scepticism*, Puhl, K., ed., 170-188. De Gruyter, Berlin.

Macdonald, C., 1989: *Mind-Body Identity Theories*. Routledge.

Macdonald, C., 1990: Weak Externalism and Mind-Body Identity, *Mind*, Vol.99, 395: 387-404.

Macdonald, C., 1995: Psychophysical Supervenience, Dependency, and Reduction, in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 140-157. Cambridge University Press.

MacMurray, J., 1970 (1961): *Persons in Relation, Vol. II of The Form of the Personal*. Faber and Faber.

Madell, G., 1988a: *Mind and Materialism.* Edinburgh University Press.

Madell, G., 1988b: Physicalism and the Content of Thought, *Inquiry*, 32: 107-121.

Malcolm, N., 1977 (1972): Thoughtless Brutes, in his *Thought and Knowledge*, 40-57. Cornell University Press.

Malcolm, N., 1995 (1989): Wittgenstein on Language and Rules, in *Wittgensteinian Themes: Essays 1978-1989*, Von Wright, G.H., ed., 145-171. Cornell University Press.

McClintock, A., 1995: *The Convergence of Machine and Human Nature*. Avebury.

McDowell, J., 1988 (1982): Criteria, Defeasibility, and Knowledge, in *Perceptual Knowledge*, Dancy, J., ed., 209-219. Oxford University Press.

McDowell, J., 1986: Singular Thought and the Extent of Inner Space, in *Subject, Thought and Context*, Pettit, P., and McDowell, J., eds., 137-168. Oxford University Press.

McDowell, J., 1989: One Strand in the Private Language Argument, in *Grazer Philosophische Studien* 33/34: 285-303.

McDowell, J., 1991: Intentionality and Interiority in Wittgenstein: Comment on Crispin Wright, in *Meaning Scepticism*, Puhl, K., ed., 148-169. De Gruyter, Berlin.

McDowell, J., 1996 (1994): *Mind and World.* Harvard University Press.

McGinn, C., 1982: The Structure of Content, in *Thought and Object: Essays on Intentionality*, Woodfield, A., ed., 207-258. Oxford University Press.

McGinn, M., 1998: The Real Problem of Others: Cavell, Merleau-Ponty and Wittgenstein on Scepticism about Other Minds, *European Journal of Philosophy*, Vol.6, No.1, 45-58.

McLaughlan, B., 1995: Varieties of Supervenience, in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 16-59. Cambridge University Press.

Moser, P.K., and Trout, J.D., 1995: Physicalism, Supervenience and Dependence, in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 187-217. Cambridge University Press.

Murdoch, I., 1992: *Metaphysics as a Guide To Morals*. Penguin Books.

Noonan, H.W., 1993: Object-Dependent Thoughts: A case of Superficial Necessity but Deep Contingency? in *Mental Causation*, Heil, J., and Mele, A., eds., 283-308. Oxford University Press.

Olafson, F., 1995: *What Is A Human Being*? Cambridge University Press.

Poland, J., 1994: *Physicalism: The Philosophical Foundations*. Oxford University Press.

Post, J.F., 1995: "Global" Supervenient Determination: Too Permissive? in *Supervenience: New Essays*, Savellos, E.E., and Yalçin, U.D., eds., 73-100. Cambridge University Press.

Putnam, H., 1975a (1973): Philosophy and our Mental Life, in his *Mind, Language and Reality: Philosophical Papers Volume 2*, 291-303. Cambridge University Press.

Putnam, H., 1975b (1967): The Nature of Mental States, in his *Mind, Language and Reality: Philosophical Papers Volume 2*, 429-440. Cambridge University Press.

Putnam, H., 1975c: The Meaning of 'Meaning', in his *Mind, Language and Reality: Philosophical Papers Volume 2*, 215-271. Cambridge University Press.

Scheer, R.K., 1991: Thinking and Working, *Philosophical Investigations*, 14, 4: 293-310.

Schroeder, S., 1995: Is Thinking a Kind Of Speaking? *Philosophical Investigations*, 18, 2: 139-150.

Searle, J., 1983: *Intentionality*. Cambridge University Press.

Searle, J., 1991 (1984): *Minds, Brains & Science*. Penguin Books.

Smith, A.D., 1993: Non-Reductive Physicalism? in *Objections to Physicalism*, Robinson, H., ed., 225-250. Oxford University Press.

Snowdon, P., 1980-1: Perception, Vision, And Causation, *Proceedings of the Aristotelian Society*, 81: 175-192.

Strawson, P.F., 1979: Perception and its Objects, in *Perception and Identity: Essays Presented to A.J.Ayer*, McDonald, G., ed., 41-60. London: Macmillan.

Van Gulick, R., 1993: Who's In Charge Here? And Who's Doing All The Work? in *Mental Causation*, Heil, J., and Mele, A., eds., 233-256. Oxford University Press.

Williams, M., 1991: Blind Obediance: Rules, Community and the Individual, in *Meaning Scepticism*, Puhl, K., ed., 93-125. De Gruyter, Berlin.

Winch, P., 1987 (1980-1): Eine Einstellung Zur Seele, in his *Trying To Make Sense*, 140-153. Basil Blackwell.

Wittgenstein, L., 1967 (1953): *Philosophical Investigations*, Anscombe, G.E.M., trans; Anscombe, G.E.M., Rhees, R., and Von Wright, G.H., eds. Basil Blackwell.

Wittgenstein, L., 1979 (1969): *On Certainty*, Paul, D., and Anscombe, G.E.M., trans; Anscombe, G.E.M., and Von Wright, G.H., eds. Basil Blackwell.

Wittgenstein, L., 1981 (1967): *Zettel*, Anscombe, G.E.M., trans; Anscombe, G.E.M., and Von Wright, G.H., eds. Basil Blackwell.

Wittgenstein, L., 1980a: *Remarks on the Philosophy of Psychology, Vol. I*, Anscombe, G.E.M., trans; Anscombe, G.E.M., and Von Wright, G.H., eds. Basil Blackwell.

Wittgenstein, L., 1980b: *Remarks on the Philosophy of Psychology, Vol. II*, Luckhardt, C.G., and Aue, M.A.E., trans; Von Wright, G.H., and Nyman, H., eds. Basil Blackwell.

Wittgenstein, L., 1992: *Last Writings on the Philosophy of Psychology, Vol. II: The Inner and the Outer*, Luckhardt, C.G., and Aue, M.A.E., trans; Von Wright, G.H., and Nyman, H., eds. Basil Blackwell.

Witmer, D.G., 1998: What is Wrong With The Manifestability Argument For Supervenience, *Australian Journal of Philosophy*, Vol.76, No.1, 84-89.

Wolgast, E., 1998: Mental Causes and the Will, *Philosophical Investigations*, 21, 1: 24-43.